智算平台运维运营技术研究报告

复旦大学
中国信息通信研究院云计算与大数据研究所
阿里云计算有限公司

2024年11月

编委会 / EDITOR

编委（排名不分先后）：

复旦大学：

吴力波、漆远、颜波、程远、韩丽妹、孙祥、张泰玮、李孟渚、张凯、葛治文、吴悠、关惠宇、黄岳、郭昕、蒋晨、徐跃东、林长龙、侯帅、江润丰

中国信息通信研究院云计算与大数据研究所：

栗蔚、马飞、苏越、赵伟博、桑柳

阿里云计算有限公司：

孙磊、付来文、李冬青、周昌盛、刘恩奇、王威、曹玉嘉、郎翊宇、杨仁远、张圣良

参编单位：

复旦大学

中国信息通信研究院云计算与大数据研究所

阿里云计算有限公司

版权声明 / Copyrig Notice

目录 / CONTENTS

前言 / FOREWORD

在数字化转型的浪潮中，智算中心扮演着越来越重要的角色，在国家数字经济和科技创新战略中的地位日益凸显。随着算力需求的不断攀升，智算中心不仅成为支撑人工智能、大数据、云计算等前沿技术发展的基石，更是推动经济社会发展的关键力量。

智算平台的运维运营是确保其高质量、稳定运行的关键。本研究报告基于复旦大学CFFF（Computing for the Future at Fudan）和阿里云智算中心的建设、运维、运营经验及中国信息通信研究院在此领域的研究成果，构建智算平台运维运营框架及评价体系。智算平台运维运营主要由三大能力域构成，一是AI运营，致力于人工智能模型的全生命周期管理，二是平台运营，着眼于提升用户体验和资源管理效率，三是平台运维，通过管理算力设备保障智算平台的业务连续性和系统安全。为客观衡量智算平台的运维运营水平，本报告从智算平台的基础设施、AI运营、平台运营和平台运维四个能力维度展开研究，提取通用、专用评估指标，构建智算平台运维运营评价体系，以期为行业内智算平台的建设、运维运营、能力评价提供参考。

智算平台运维运营是一个充满挑战的新兴领域，需要不断探索和创新。本研究报告旨在为业界提供更加全面、深入的研究视角，以促进智算平台运维运营的专业化、标准化和智能化发展。本研究报告仍有不足指出，期待业界专家和广大读者提出宝贵的意见和建议，共同推动智算平台运维运营领域的发展与完善。

01

研究背景及价值

算力的现状和发展趋势

智算平台的现状和发展趋势

智算平台的运维运营现状与面临挑战

研究背景及价值

1.1 算力的现状和发展趋势

随着数字化转型的深入和人工智能、大数据、云计算等新兴技术的广泛应用，算力已成为支撑经济社会发展的关键基础设施。中国作为全球第二大经济体和数字技术应用的前沿阵地，其算力需求呈现出爆发式增长态势。2024年政府工作报告中提出，大力推进现代化产业体系建设，加快发展新质生产力。要深入推进数字经济创新发展，制定支持数字经济高质量发展政策，积极推进数字产业化、产业数字化，促进数字技术和实体经济深度融合。深化大数据、人工智能等研发应用，开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。实施制造业数字化转型行动，加快工业互联网规模化应用，推进服务业数字化，建设智慧城市、数字乡村。深入开展中小企业数字化赋能专项行动。支持平台企业在促进创新、增加就业、国际竞争中大显身手。健全数据基础制度，大力推动数据开发开放和流通使用。适度超前建设数字基础设施，加快形成全国一体化算力体系。我们要以广泛深刻的数字变革，赋能经济发展、丰富人民生活、提升社会治理现代化水平。

中国算力的快速发展为数字经济提供了强有力的支撑。随着“东数西算”工程的推进，中国的算力布局更加优化，特别是智能算力的快速增长，为中国在AI和大数据时代的增长提供基础。未来，中国将继续加强算力基础设施的建设，推动技术创新，完善政策和标准体系，构建全产业链生态，以促进算力产业的健康、高效和可持续发展。

1.2 智算平台的现状和发展趋势

本研究报告讨论的智算平台，是指通过使用大规模异构算力资源，用智能算力(GPU、FPGA、ASIC等)，主要为人工智能应用(如人工智能深度学习模型开发、模型训练和模型推理等场景)提供所需算力、数据和算法的设施。智算平台作为算力产业的重要设施，支撑着人工智能及相关产业的快速发展，但当前的智算平台多采用硬件驱动模式，存在水平较低、分割化严重、生态建设不足等问题，难以满足AI对“大数据、大计算、大模型”的需求。当前，美国等国家在算力、算法和数据方面已形成先发优势，而中国的公共智算平台及生态与之存在差距，特别是在AI公共算力设施及部分AI芯片上。AI算力及其服务市场可能出现“碎片化”，低水平、小规模的智算中心无法支撑大模型训练任务，可能导致资源浪费。面对上述等形势，国家和地方政府积极出台相关政策，推动智算平台的建设和算力产业的发展。

为了支持通用AI的发展，满足不同场景下的算力需求。智算平台将弥补传统计算中心的局限性，提供更广泛的服务，满足更多行业和领域的算力需求。此外，智算平台也通过优化算力资源配置、支持实时和离线计算需求等方式节约能源和成本。未来一段时间里，高性能算力产业生态体系也会是建设重点，推动产业链上下游协同发展，形成统一开放的AI算力产业生态。智算平台的发展可以降低中小企业的算力使用门槛，提升算力设施的普惠服务能力，加速赋能各行各业，推动产业数字化转型。智算平台正处于快速发展阶段，未来智算平台的建设也会是算力建设的重点，为算力的蓬勃发展提供平台服务基础。

智算平台离不开算力相关服务的专业化。算力相关服务作为智算平台的核心支柱，其发展必须与智算平台的整体进步相匹配，以确保整个系统的协调性和效能。算力服务的专业化不仅体现在技术层面的深度融合和智能化管理，更在于服务模式的创新、生态构建的完善以及安全合规的强化：通过结合LLMOps等思想，实现算力资源的智能调度和优化配置，提升服务效率和响应速度；探索按需服务、弹性服务等新型服务模式，以满足用户在多样化和个性化算力需求方面的期望，增强服务的灵活性和适应性；构建开放、共享的算力服务生态系统，促进跨行业、跨领域的协同创新和资源共享，以实现算力服务的可持续发展；加强算力服务的安全性和合规性，确保数据安全和用户隐私得到有效保护，构建用户信任的基石。

智算平台运维运营的价值

智算中心投资规模巨大，其能力与运营效率将成为运作的关键，构建合适的运维运营体系可有效地保持智算平台长期稳定运行，高效地管好和用好算力，并提供管理的实践，技术和工具的集合。

智算平台的运维围绕着模型服务，算力服务，容器服务，网络服务，存储服务以及安全服务等方面进行。智算平台的运营包含用户的日常管理及 AI 运营两个重点，用户运营包括用户管理、用户答疑、账单收费、工单管理、知识库建设等方面，AI 运营包括数据集运营、模型管理和部署、模型微调、提示词工程等能力。

智算运维运营平台为工程师提供了一个协作环境，该环境促进了数据和模型迭代探索、实时协作实验跟踪、提示词工程以及模型 Pipeline 的管理。同时，它还支持对大型语言模型（LLM）的控制模型转换、部署和监控。整体方案提供了一套完整的AI生命周期管理服务，从开发到部署再到维护，确保了平台的高效运行和持续优化。建设智算运维运营平台和相关团队，可以为平台带来如下保障：

1.确保服务连续性：
通过有效地运维运营，智算平台能够保证服务的连续性和稳定性，避免因故障或性能问题导致的服务中断，通过日常巡检和监控可以降低重大故障的发生概率。

2.提升用户体验：
良好的运维运营能够快速响应用户需求，提供及时的技术支持和问题解决方案，从而提升用户满意度。

3.研发效率提升：
通过工具研发的支持，智算运维运营平台允许团队更快地开发模型，提供更高质量的模型，并更快地部署到生产环境中。

4.优化资源利用：
通过精细化的资源管理和调度，可以提高计算资源的利用率，避免资源浪费，降低运营成本。

5.知识管理：
建设和维护知识库，促进使用方法和经验的共享，降低初学者的门槛。

6.模型微调、推理和监控：
模型微调优化模型以执行特定于领域的任务。模型推理可以基于现有知识管理生成内容。

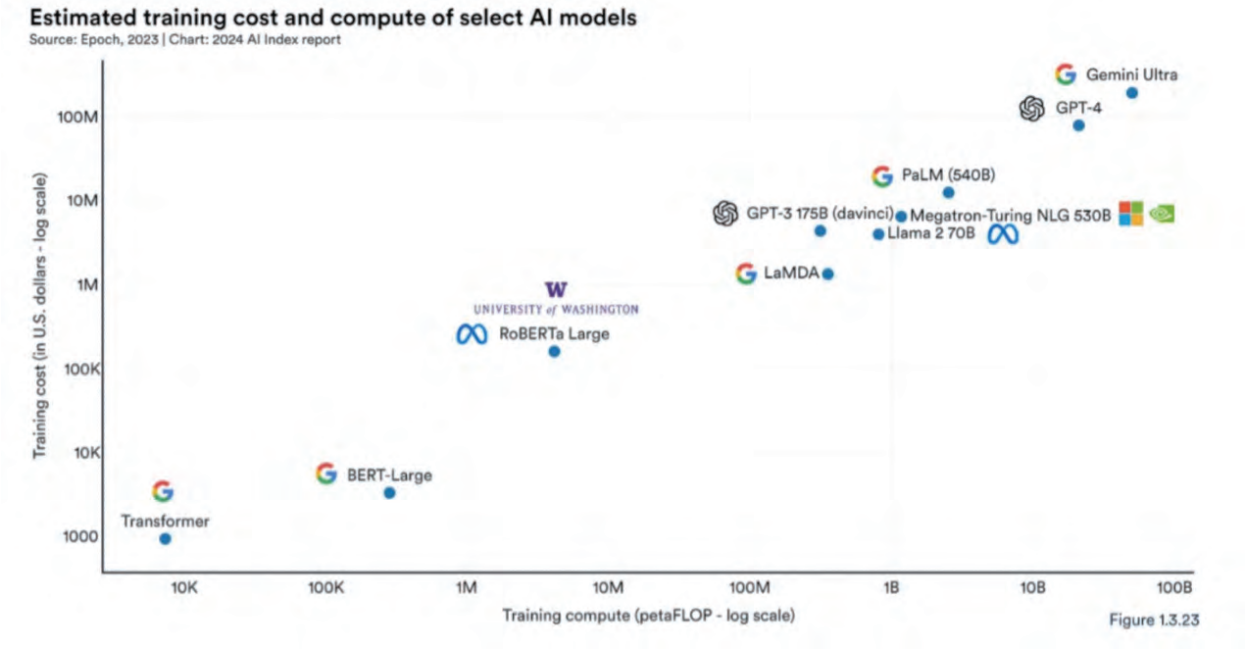
7.确保模型性能：
通过持续的监控和维护，智算运维运营可以确保模型在生产环境中的性能稳定，及时调整以适应新的数据和需求。

8.可扩展性：
随着业务需求的增长，智算运维运营支持平台的无缝扩展，可以灵活地增加计算和存储资源。

复杂的 AI 技术和智算生态对用户是一个挑战。高质量的智算平台的运营运维能力不光可以提升平台的稳定性，做好资源和用户管理，同时也降低 AI 模型的研发门槛，将研发好的 AI 模型快速应用到实际场景中。尤其对于那些工程能力相对薄弱的组织，如部分中小企业、具有 IT 诉求的非 IT 企业，智算平台的运维运营能力尤为关键。这些组织可能缺乏独立维护复杂 AI 平台的经验，依赖外部提供的高质量运维运营服务，可以加速创新孵化过程。

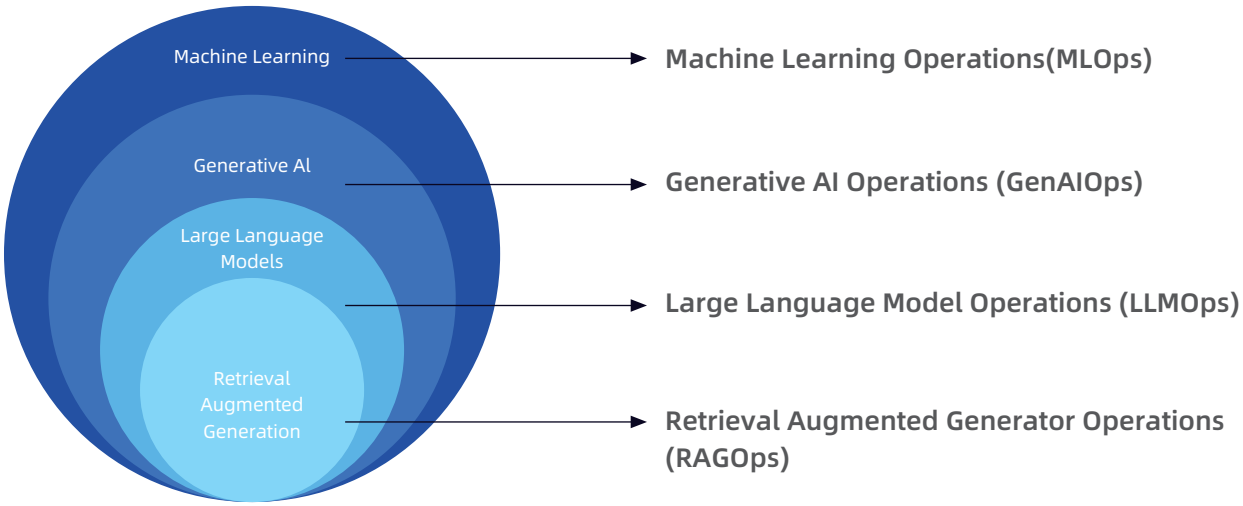
1.3 智算平台的运维运营现状与面临挑战

随着 AI 技术的发展，算力训练需求增长，智算设备紧缺，训练大型 AI 模型的成本变得极其高昂。OpenAI 的 CEO Sam Altman 曾透露，GPT-4 模型的训练成本超过了1亿美元。



▲ 图 1 大模型训练算力需求变化

国外在智算平台的建设和运维方面积累了丰富的技术和实践经验，有专业的团队负责智算平台的建设和运维工作。这些团队通常具备跨学科的知识和技能。目前，已经出现了 LLMOps 的概念，除了计算资源的管理和调度之外，还包括对 AI 模型全生命周期的管理能力。



▲ 图 2 AI领域不同层级运维理念

当前，国内智算平台运维运营相关领域的资料有限，尚未形成体系化的智算平台运维和运营解决方案。智算平台运维运营方面的不足，以及完善运维运营体系的必要性，主要体现在以下几个方面：

1.缺乏成熟的运维运营模式和经验。

2.缺少交叉学科的专业人才和团队。

3.需要建立更加完善的算力资源管理和调度机制。

4.运维运营缺乏对AI模型全生命周期管理的深入理解和实践。

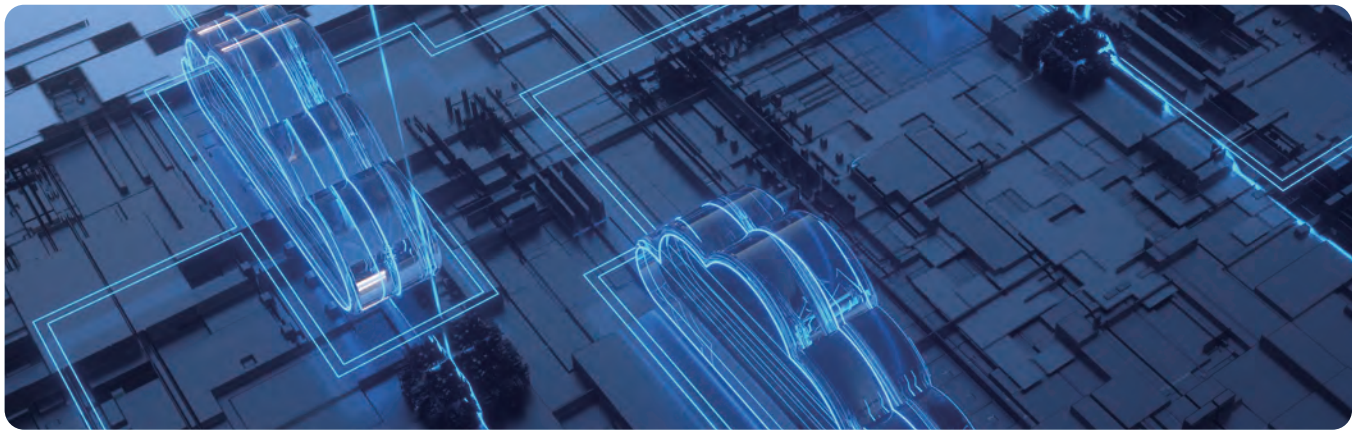
5.需要加强智算平台的安全性和稳定性。

随着用户的增加、算力供给增长以及服务生态的多样化，智算平台的运维和运营存在着较大的挑战，主要体现在人才缺失、流程和工具化能力缺乏、相关技术门槛高运营运维难度大、任务失败后排障困难等几个方面：

1.3.1 人才供给挑战

人才供给挑战主要体现在两方面，一是人才紧缺，缺乏具备必要专业知识和技能的人才，导致招聘难度增加；二是传统运维难度大，传统运维方式面临挑战和巨大的学习成本，缺乏高效的运维经验和标准。

传统运维运营方法与智算平台的运维运营要求之间存在较大差距，主要体现在AI模块的运维支持上。人工智能技术作为近几年的新兴的领域，综合了机器学习、深度学习、自然语言处理和计算机视觉等技术，对问题排查的人员能力要求很高，目前都由全栈型和有经验的算法工程师解决。



知识领域	传统运维运营人员	智算平台运维运营人员	备注
算力规模与性能			
CPU集群算力	1	1	集群传统高性能计算任务，需要运维机器
GPU集群算力	0	1	AI任务主要为各种GPU卡，需要运维机器和升级驱动
高性能存储	0	1	高性能存储，吞吐快，性能提升
平台服务能力			
预装软件数量	0	1	科研软件安装，外部工具接入
交互式环境	0	1	算法开发环境的使用和DEBUG
性能优化	0	1	AI任务和资源优化
任务调度	0	1	诊断K8s和Slurm调度的问题
运维与运营			
运维监控	1	1	帮助用户建设机器的运维体系，并且进行平台的变更操作
规章制度	1	1	帮助用户建机器使用制度
运营团队	0	1	协同机器的运营和平台的管理
AI和高性能计算			
HPC任务	0	1	HPC任务的使用
AI建模	0	1	模型的训练和推理
大语言模型框架	0	1	并行计算，PyTorch 和 TensorFlow 等框架
大数据	0	1	大数据加工，数据传输
算法开发与服务	0	1	协同用户解决算法训练的运维问题
AI和HPC镜像安装管理	0	1	为用户下载和安装镜像
其他			
网络性能	1	1	网络问题的诊断
容器化技术	1	1	POD的诊断，重启和删除
网络安全	1	1	网络安全能力建设

▲ 表1传统运维运营人员与智算平台运维运营人员能力对比　　注：1代表运维运营人员必备能力，0代表运维运营人员非必备能力

根据斯坦福大学 2024 年发布的《Artificial Intelligence Index Report》，智算领域的技术进步正在加速，但同时也带来了对专业人才和先进设备的巨大需求。而能够理解 AI 又愿意做智算运维运营平台的人员非常稀缺，招聘难度巨大。

Top 10 specialized skills in 2023 AI job postings in the United States， 2011-13 vs. 2023

Source： Lightcast，2023Chart： 2024 AI Index report

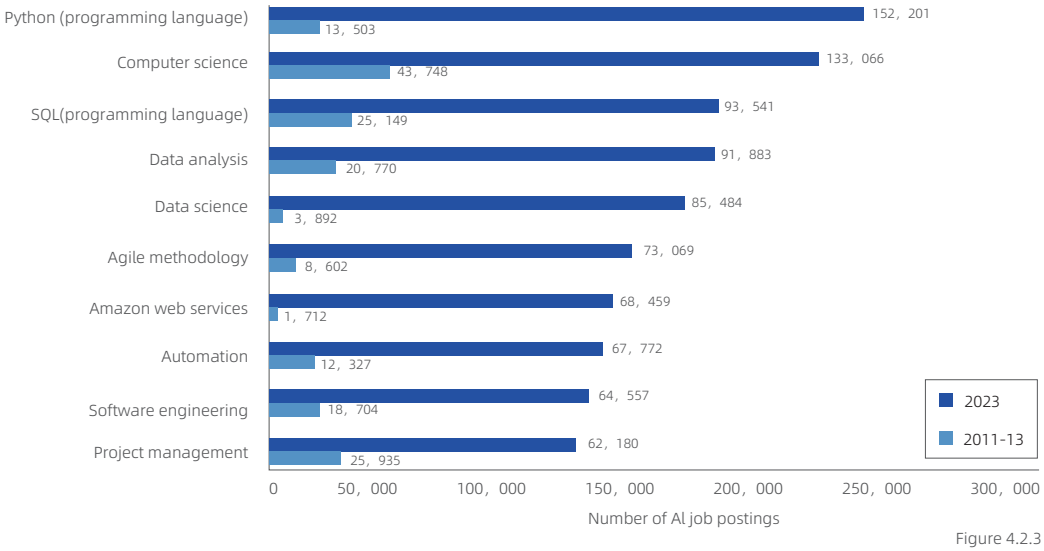


Figure 4.2.3

▲ 图 3 2023年美国AI从业人员top10工作技能具备人数与十年前对比

随着技术人才使用的智算平台设备日益昂贵，对运维运营人员的要求也相应提高。不仅要求他们掌握高水平的专业技术能力，更要具备出色的管理与决策技能，以保障智算平台的高效运行和持续创新。当前国内在这一领域面临运维运营人才短缺的问题，亟需在持续的教育和实践中培养。合适的智算平台运维运营人员，不仅要有传统运维运营的基础，还要对人工智能技术有深刻理解，掌握相关的管理和决策知识，以适应智算平台在数字化转型和AI升级中的新需求。

1.3.2 流程和工具化能力缺乏

目前，大模型训练的生态系统仍在建设之中，相关的流程和工具尚未完全产品化。同时，我们还缺乏统一的标准和接口来管理相关资源。例如，对于模型的运行状态、对应的 GPU 机器以及平台稳定性，我们还需要一个统一的监控和统计系统；大规模 GPU 集群的扫描软件、AI 训练生态系统，推理和模型输出等都处于创新阶段。偏定制化的需求，面临流程缺失和工具缺乏等问题，极大程度地增加了运维运营工作的难度，目前市面上类似 Datadog、HuggingFace、atabricks 等公司都在积极地解决 AI 任务监控和训练的生态问题，未来有望可以标准化输出。

1.3.3 智算门槛高，运营运维难度大

目前智算的高技术门槛和运营运维的复杂性使得许多企业和研究机构望而却步，其主要原因在于对 GPU 资源的大规模依赖。此外，智算系统的设计和实现需要跨学科的知识和技能，包括机器学习、数据科学、软件工程等，均成为了运维运营工作开展的挑战。

在运营和维护智算系统时，团队面临的挑战尤为严峻。系统稳定性的维护需要持续地监控和及时故障排除，而性能优化则要求对系统架构有深入的理解。随着技术的快速发展，智算系统需要不断地更新和升级，以适应更大规模的算法参数和更大的数据集,通过更敏捷的模型应用部署平台，来满足 AI 模型对实际业务场景的适配。为了克服这些挑战，企业和研究机构需要投入更多的资源进行人才培养、技术研发，并探索和总结更高效的运维和运营策略。

1.3.4 任务失败后排障困难

计算任务失败原因分析路线非常复杂，从硬件到上层框架链路长，涉及的领域众多，对目前运维运营人员的技术要求较高。任务排障困难体现为如下几方面：

1.系统架构复杂：
智算平台通常由多个模块组成，如底层基础设施、机器学习平台和运维运营平台等，每个模块都有其特定的功能和架构，问题定位困难。

2.硬件和软件问题：
底层硬件问题（如ECC错误、NVLink错误）和软件配置问题（如Shell启动失败、缺少配置文件）可能影响系统运行，需要专业知识进行诊断。任务调度失败、训练速度慢、资源不足（如OOM错误）等问题，需要对平台执行 AI 任务的逻辑有一定了解。

3.用户权限和资源管理：
用户权限设置、资源申请、工作空间配置等方面的问题，需要对平台的运营体系有深入了解才能解决。

4.环境配置和依赖问题：
AI 模型训练环境配置复杂，涉及镜像、数据集、代码等 AI 资产的管理，以及依赖包的安装和配置问题。

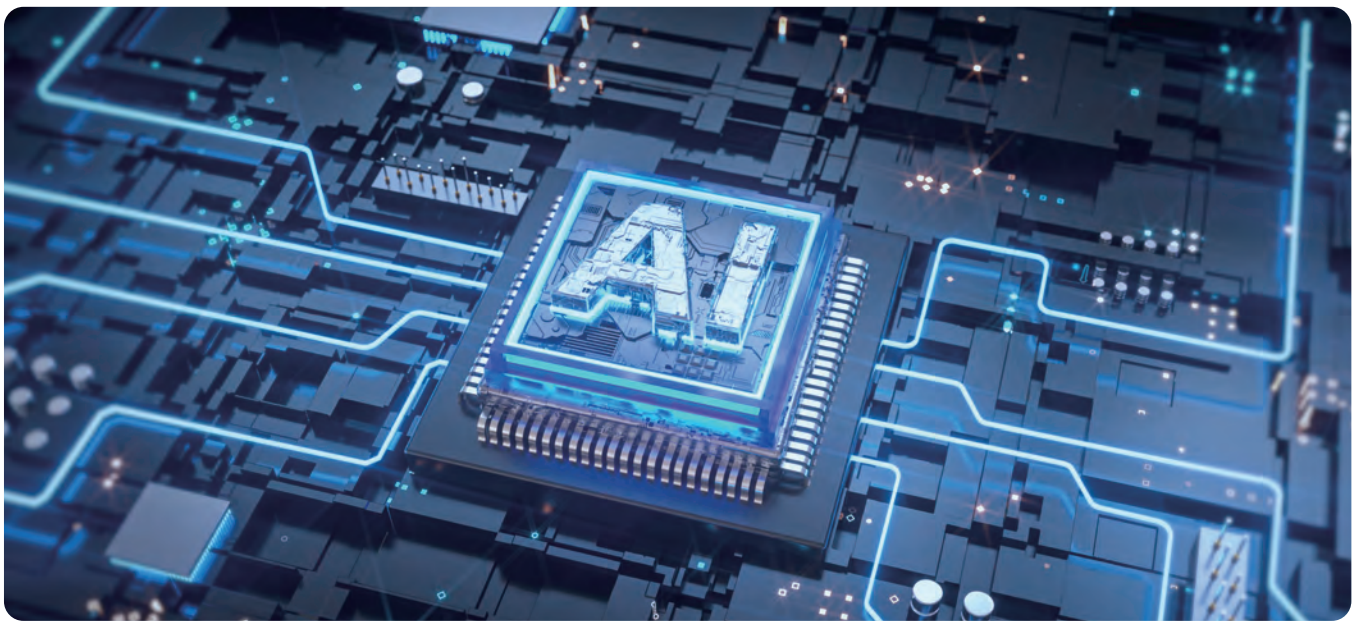
5.网络和存储问题：
网络连接问题、存储设置错误、文件操作限制等，均可能影响用户的正常使用。

6.硬件故障：
GPU 卡顿、掉卡、硬件损坏等问题，需要硬件维护团队及时介入。

7.用户熟悉度不足：
用户对平台的使用不熟悉，导致操作错误或无法充分利用平台功能。

智算平台的任务排查是一项极具挑战的工作，它要求运维人员不仅要有深厚的技术背景，还需对整个系统架构有全面的理解。从底层硬件的稳定性到软件配置的精确性，每一个环节都可能导致训练任务执行失败。同时新的挑战不断涌现，如确保数据安全、遵守合规性要求、处理大规模并发请求等，都进一步增加了任务排查的难度。

根据目前智算平台运维运营的现状，为了提高智算平台的运维效率和稳定性，需要完善自动化监控和故障排除工具，加强人才培养，确保智算平台在面对日益复杂的AI任务时，仍能保持高效和稳定，并且将大模型等AI技术有效得应用。本研究报告面向智算平台支持AI模型训练的全生命周期，总结当前智算平台的运维和运营难点，并提出了相应的解决方案。



02

智算平台运维运营

智算平台运维运营中心主要功能

智算平台运维运营组织架构及制度体系

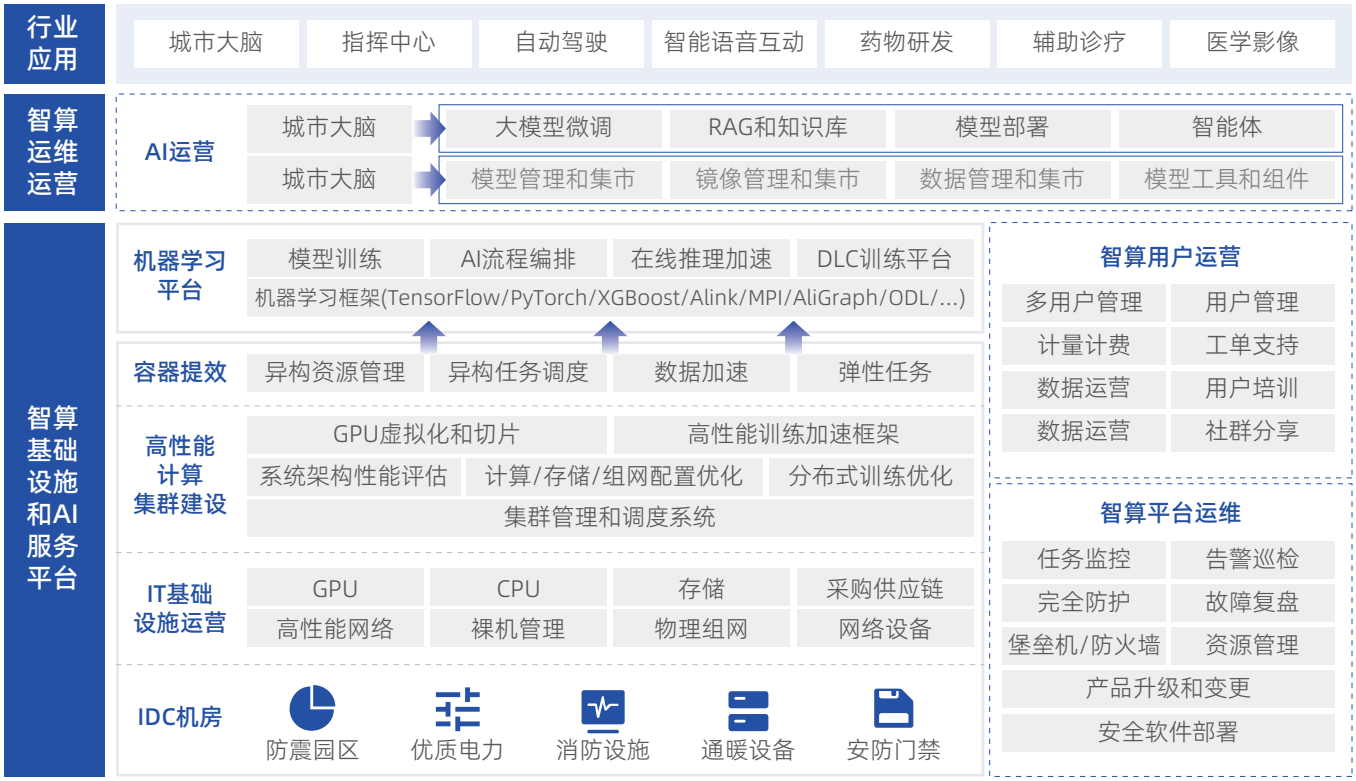
AI运营

智算平台运营

智算平台运维

2. 智算平台运维运营

智算平台为支持多样的行业应用而建设，智算平台运维运营在智算平台体系结构中的位置如下图所示：



▲ 图 4 智算平台运维运营体系结构

智算基础设施和AI服务平台位于智算平台体系结构的最底层，主要提供两个重点能力：基础设施 IaaS 和 AI 平台 PaaS。用户无需组建和运维复杂的 GPU 机器、存储和 RoCE 网络，即可使用高拓展性、高性能的 IaaS+PaaS 的环境：

1.基础设施IaaS：

IDC 机房、网络交换机（RDMA网络交换机、通用网络交换机）、算力服务器（智算算力服务器、通用计算服务器）、存储服务器等能力；同时还有基于基础设施的集群建设，为上层平台和应用提供算力、存储、网络、容器、容器镜像、安全等服务。

2.AI平台PaaS：

提供随开即用的的 AI 作业平台，完成对 AI 模型（包括大模型）的开发和训练。

智算运维运营在智算基础设施服务平台和行业 AI 场景之间起到桥梁作用，除了为用户提供统一的资源管理和算力资源的监控，也为上层的智算模型运营提供产品和服务（模型微调、Agent、AI资产生态运营等），有效地提升智算平台整体的性能和用户体验。

通常智算运维运营能力由智算平台运维运营中心承载，其能力进一步细分为 AI 运营、平台运营及平台运维。

2.1 智算平台运维运营中心

▲ 图 5 智算中心运维运营总览

智算运营运维中心主要分为三个重点的模块：

2.1.1 AI运营

AI 模型的开发，尤其是大语言模型的开发过程包含许多复杂组件，如数据加工、数据预处理、提示词工程、模型微调、模型部署、模型监控等，同时还需要跨团队的协作和交接，从数据工程到数据科学再到机器学习工程，整体流程需要严谨运营以及相对应的产品工具。AI 运营的目标是通过产品工具和专家服务，降低用户在 AI 模型训练和应用的工程门槛，提高大模型应用的开发效率。

AI 运营非常重要，包括可视化，透明度和可解释性。通过AI模型的运营模块，可以让非技术人员参与到 AI 应用的运作中。其中主要包含模型运营和 AI 资产运营。

1.模型运营：

模型运营的目标是为了释放大模型的价值，其中包含模型微调、提示词工程、智能体（包含各种工程组件）以及模型监控等能力，同时也包含大模型专家服务用来解决在模型训练和推理过程中遇到的问题。

2.AI资产运营：

主要面向丰富的 AI 资产生态，具体包含：

1) **模型集市：**包含官方开源的大模型和组织内公开的大模型，可以进行模型版本的控制更新，分享和部署。
2) **数据集市：**包含官方开源的数据集，和组织内公开的数据集，可以协同开展数据上云，数据加工，数据共享等。
3) **镜像集市：**主要包含支持各种大模型的不同镜像，来自不同的社区。
4) **实验集市：**主要包含各种业务组件，用于降低模型部署或者数据加工的工程化门槛。

2.1.2 平台运营

平台运营可以帮助企业利用已有的算力资产，向租户出售算力产品和增值服务，帮助用户更高效地使用算力。同时平台运营会有效地处理用户资源数据，给企业组织提供决策和实现平台，从而提高整体智算平台的运营效率，降低管理和维护成本。

1.用户运营：

用户运营主要包含用户权限管理、工单答疑、用户培训等。通过工单服务解决用户找人难、上手难、排查难的痛点。通过智算知识库帮助管理者在运营过程中持续沉淀智算行业的宝贵经验。

2.资源运营：

一站式的资源全生命周期管理。资源运营主要包含全面的资源管理，包含不同类型、不同收费模式和计算资源进行混合管理。用户能够在平台对计算资源进行从申请、审批、创建、变更到回收的全链路管理动作，并且平台能够精确记录资源的申请或变更记录、资源的项目归属和资源的计费主体。

3.运营管理：

包含管理经验的运营流程设计、数字化管理、经营分析和计量计费等模块，帮助用户高效、便捷地对智算场景开展更全面精细和准确的运营。通过数字化管理和经营分析可以快速的发现问题，提升用户体验和资源利用率。

2.1.3 平台运维

通过端到端地对物理资源、机器学习平台及上层应用进行日志采集和监控，平台运维能够快速且精确地诊断问题，迅速响应并预防重大问题的发生。同时，平台运维提供专门针对智能计算任务的运维服务，以解决用户在使用时硬件基础设施时遇到的功能和性能等问题。

1.业务连续性：

业务连续性需要软硬一体的全栈运维来支持，覆盖了从各个型号的 GPU、CPU 硬件、并行存储节点，以及网络和通信等底层基础设施硬件。此外，还需支持上层的容器服务，确保容器和容器间的通信，以及每个容器里代码平稳地运行，从而产生可靠的 AI 运算结果。在业务连续性方面，需要全面的日志采集、业务监控、故障应急、变更升级和 AI 运维巡检能力，为整个 AI 系统提供高效地运行支持。

2.安全防护：

安全体系的设计需要多个重要的参与方，运维团队需要跟安全团队紧密合作，确保技术基础设施的可靠性和安全性。运维团队负责日常系统维护、软件部署和故障排除，安全团队则专注于评估风险、监控威胁和强化防护措施。

3.智算运维：

智算运维模块不同于传统运维的服务能力，主要针对大模型训练和推理的相关业务需求开展性能分析优化、算力和存储扩容、软件镜像安装、模型训练报错诊断、大模型迁移GPU卡等比较新兴的运维服务能力。

2.2 智算平台运维运营组织架构及制度体系

2.2.1 组织架构

为保障智算平台的安全稳定和业务的长效运营，平台运营运维需要如下组织构成。

1.智算平台运营组：

提供资源受理和办理、资源账单服务、工单受理、赋能培训、产品需求缺陷管理、解决方案服务、资源目录梳理、资源开通流程规范、资源计量计费规范、资源效能规范等服务。

2.智算运维保障组：

保障平台软硬件稳定性服务、服务完成平台日常变更、告警处理等问题处理，人员通常配置驻场运维服务 5*8，远程运维保障服务 7*24。

3.AI 应用运营组：

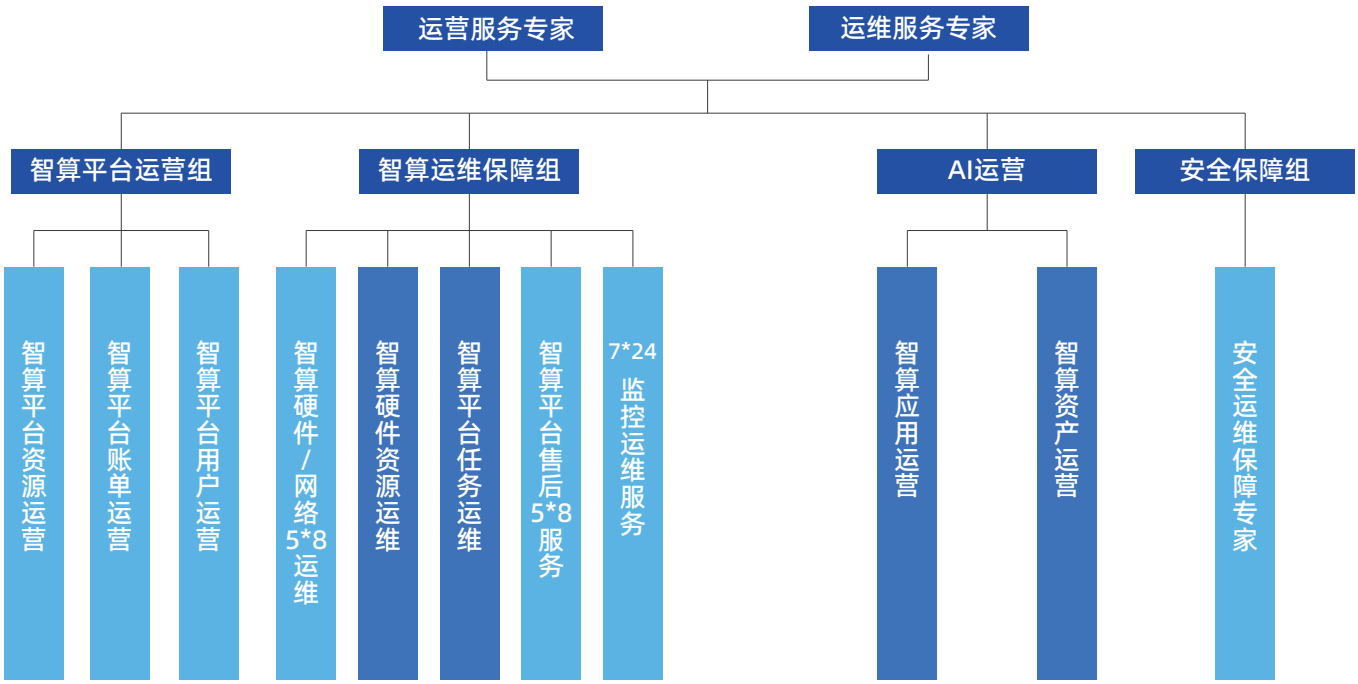
大模型模型运营负责提供模型部署的数据支持，确保模型可以稳定高效地推理和应用，同时确保用户关于 AI 模型的微调、RAG、Agent 建设可以顺利开展。

4.AI 资产运营组：

针对大模型开发过程中主要的生产资料数据集，模型和镜像进行有效的运营管理。数据运营负责收集、整理和存储用于大模型训练的数据集，确保数据的质量和完整性。同时负责为相关的模型提供镜像的下载和部署服务。

5.安全保障组：

负责平台安全架构设计，包括防御、监测体系构建，潜在风险识别和安全策略制定。



▲ 图 6 智算平台运营运维组织架构

2.2.2 制度体系

为了保障平台建设和运维运营过程中的整体稳定性和线上业务的正常运行，结合人员和工具的能力建立流程和问题管理机制。

1.资源管理：
建立资源分配和调度的规则，确保 AI 模型训练和推理任务能够高效利用计算资源。

2.故障恢复：
制定故障恢复流程，包括自动故障转移、备份和恢复机制，以最小化系统停机时间。

3.性能监控：
实施实时监控系统，跟踪集群的性能指标，如负载、响应时间、错误率等，以便及时发现并解决问题。

4.资源巡检机制：
定期进行资源巡检，确保资源配置得当，及时发现资源使用中的瓶颈和浪费问题。

5.用户管理：
建立用户管理体系，确保用户权限的合理分配，优化用户体验，包含用户在项目申请、账单结算、工单提出等多个环节的管理。

6.数据管理：
制定数据管理政策，确保数据的完整性、可用性和合规性，提高数据的质量和分析能力。

7.AI模型管理：
建立 AI 模型的全生命周期管理流程，包括模型的开发、测试、部署、监控和下线。

8.AI应用管理：
对 AI 应用进行系统化管理，确保应用的性能、安全性和用户满意度。

9.文档和知识管理：
维护详细的文档体系，记录系统架构、操作流程和故障处理案例。

10.成本管理：
监控和优化集群的运营成本，包括硬件投资、能源消耗和维护费用。

安全架构设计：

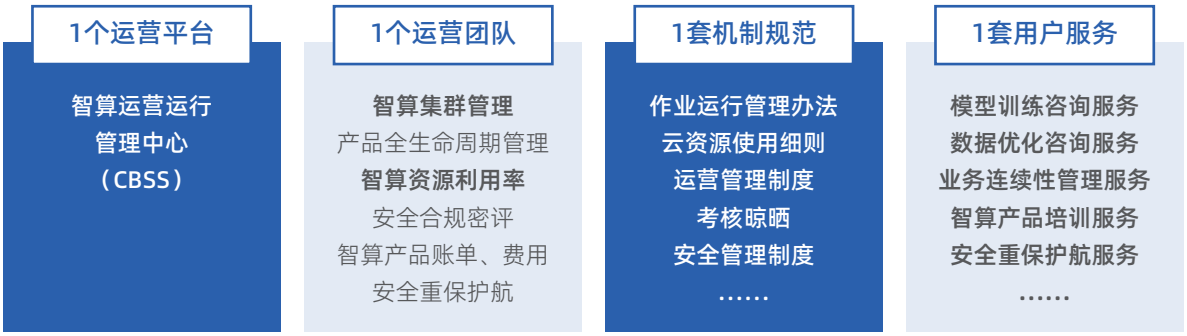
制定严格的安全政策和协议，包括访问控制、数据加密和网络安全措施，保护集群免受内外部威胁。

安全合规性和审计：

确保所有操作符合法律法规要求，并定期进行内部和外部审计。

产研协同体系：

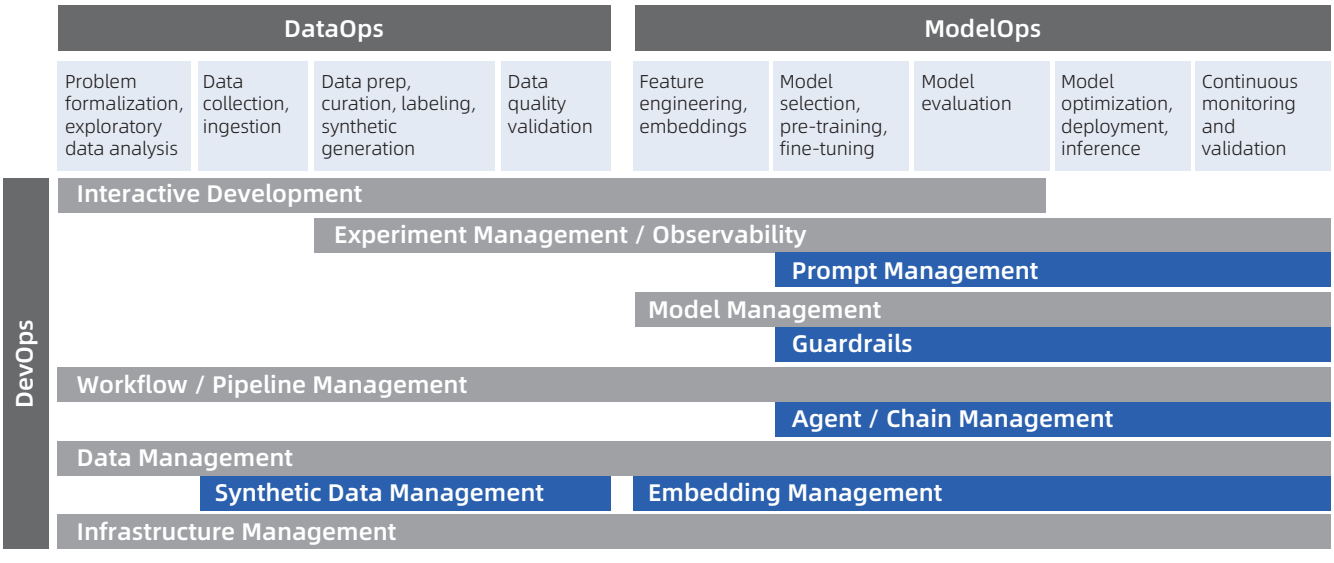
确保产品的缺陷和需求可以伴随着业务的发展快速迭代。对日常运维运营中发现的缺陷和需求可以快速解决。



▲ 图 7 智算平台运营运维制度体系

2.3 AI运营

AI 的运营主要包含模型运营和 AI 资产运营，其中模型运营主要为了完成 AI 模型的业务应用，AI 资产运营主要是为模型训练和推理提供高质量的素材。

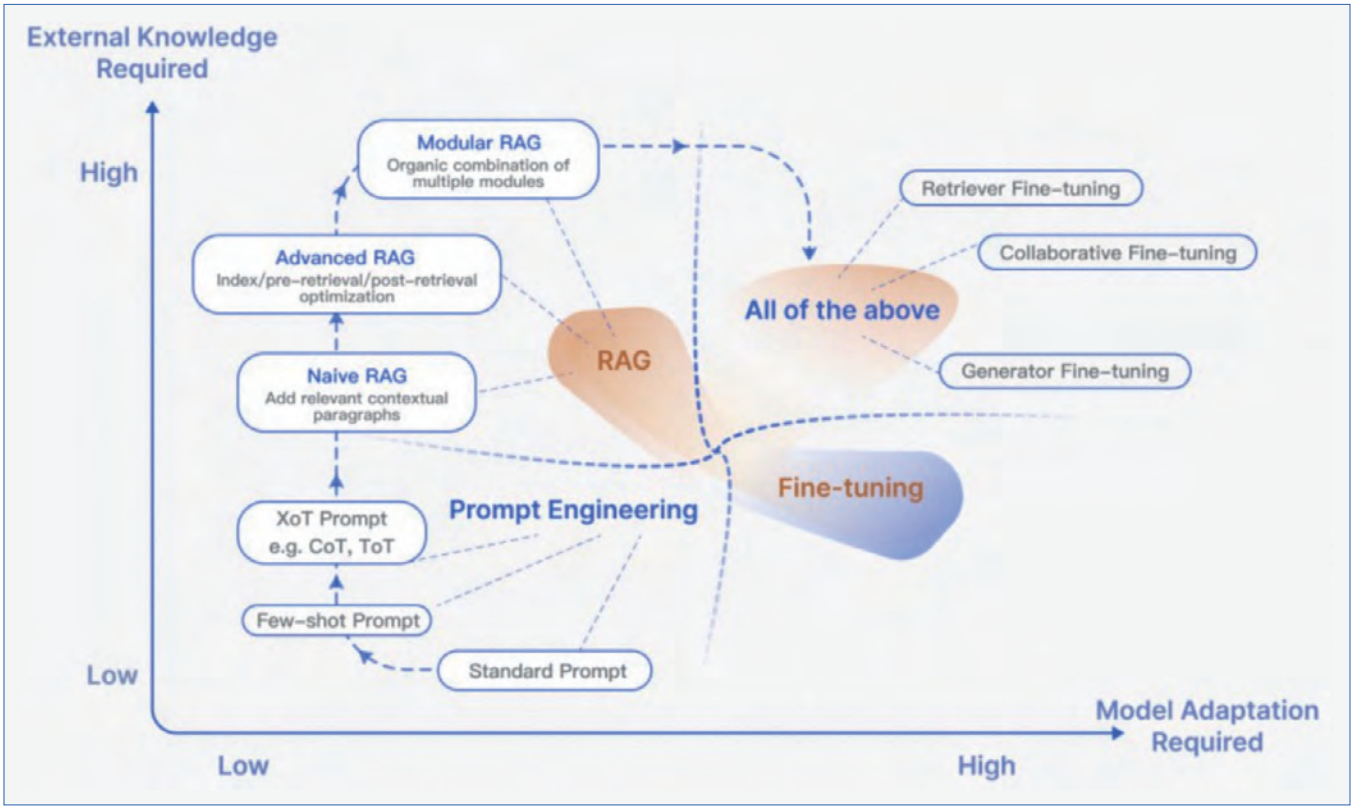


▲ 图 8 AI 运营范围划分

大模型应用还面临诸多挑战，例如开发团队还未适应大模型编程的需求，对大模型的实际应用场景理解、工具的选择（例如中间件、向量数据库等）以及团队的协作模式、如何构建 Prompt 等方面都存在一定的认知偏差。开发团队需要在大模型技术栈方面建立更多的共识，对于如何使用 RAG（Retrieval Augmented Generation）或者微调等应该有更明确的工作流程。

2.3.1 模型运营

模型运营是指通过大模型提供应用和服务。模型运营基于对外部的知识库的输入和模型是否需要调参可以分为微调、RAG 和提示词工程，帮助大模型快速回答专业领域的问题。另外智能体平台提供产品化工具，简化工程化能力，帮助用户快速部署模型和实现模型在实际业务场景中的价值。



▲ 图 9 微调、提示词工程、RAG 技术

2.3.2 模型微调 (Fine-tuning)

模型微调是指在预训练模型的基础上，针对特定的应用场景或数据集进行进一步训练的过程。可以通过预训练的 LLM 作为起点，然后在特定任务或领域的标记数据集上训练完成。这样做可以使得模型更好地适应特定的任务，提高其在该任务上的表现，其主要技术包括：

1.全微调：

用预训练模型作为初始化权重，在特定数据集上继续训练，全部参数都更新的方法。

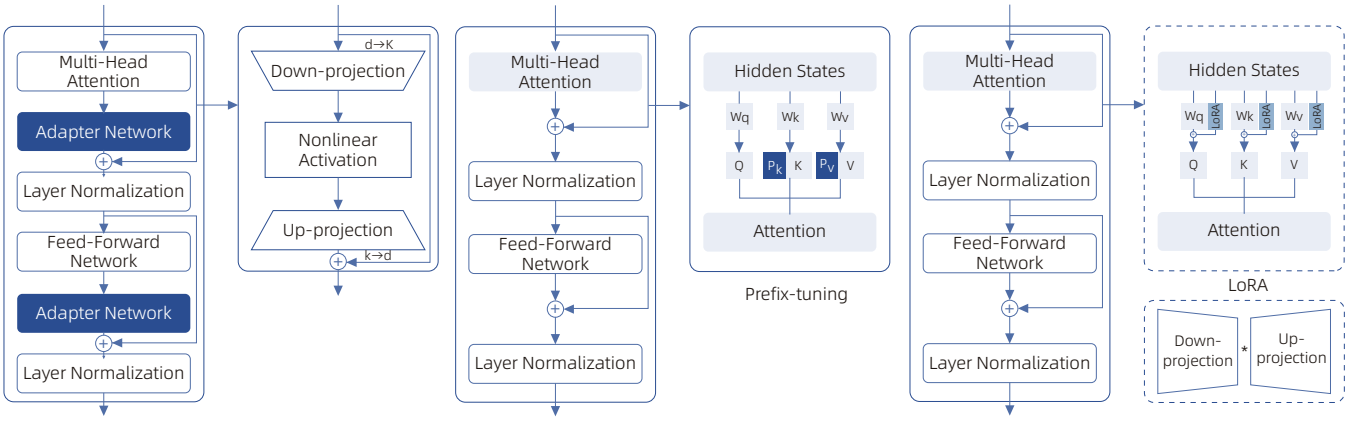
2.高效参数微调：

a) 增加额外的参数（Addition-Based）：Prefix Tuning、Prompt Tuning、Adapter Tuning。

b) 选取一部分参数的更新（Selection-Based）：BitFit。

c) 引入重参数化 Reparameterization-Based：LoRA。

d) 混合高效微调：MAM Adapter、UniPELT。



▲ 图 10 大模型参数微调原理示意图

2.3.3 RAG (Retrieval-Augmented Generation)

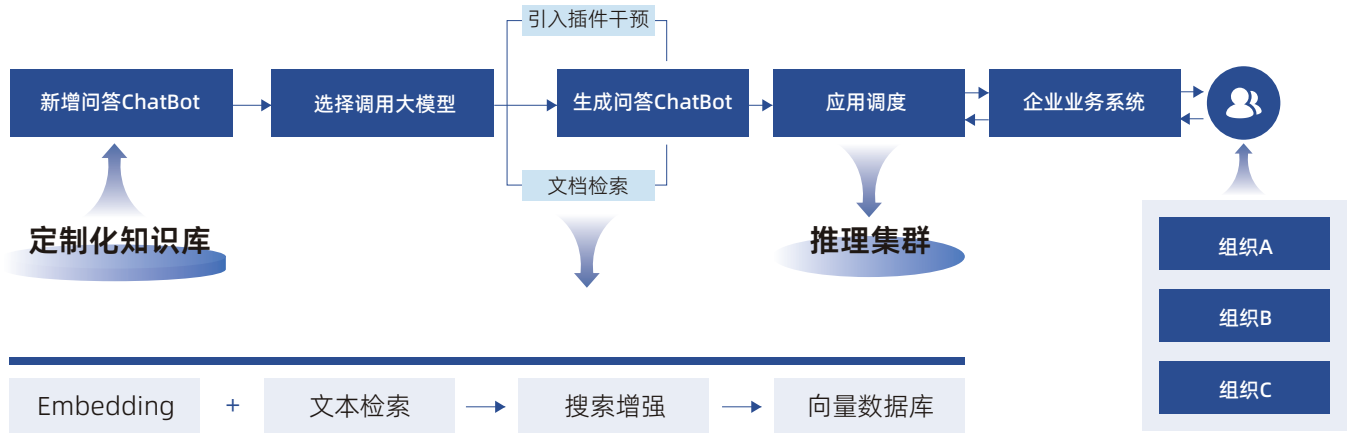
RAG 是一种结合了检索（Retrieval）和生成（Generation）的模型架构，它首先从一个大型的数据库中检索相关信息，然后将这些信息整合到生成模型中，以生成更加丰富和准确的输出。该方法有非常多的优势，例如：

a) RAG 通过将答案与外部知识联系起来，减少语言模型中的幻觉问题，并使生成的回答更加准确可靠。

b) 使用检索技术可以识别最新信息。保持了响应的及时性和准确性。

c) 透明度，通过引用来源，验证答案的准确性，增加对模型输出的可解释性。

d) 安全和隐私管理，RAG 凭借其在数据库中内置的角色和安全控制，可以更好地控制数据使用。



▲ 图 11 RAG的技术架构示意图

整体 RAG 系统包含两个阶段：检索阶段（Retrieval Phase）和生成阶段。其中在检索阶段，根据用户提出的问题，检索系统搜索用户上传的知识库，（该知识库可能包含文档、网页或其他形式的数据。同时知识库会被切成不同的片段以向量的方式存在向量库）。语言模型会把检索到的文档作为输入，结合问题和用户的原始问题，生成答案输出。

2.3.4 Prompt 提示词

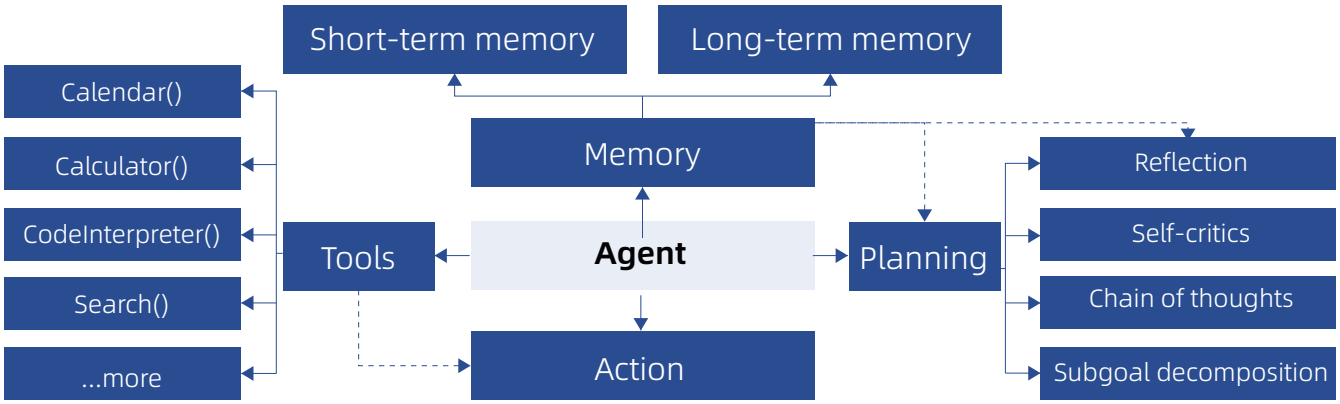
提示词是一种引导模型生成特定类型回答的方法。在一些生成模型中，通过精心设计的提示词可以引导模型生成更加相关和高质量的内容。高质量的提示词可以提升答案生成的质量，实现特定任务和目标，同时设定个性化的风格来适应多样化的需求。

2.3.5 智能体平台

智能体是人工智能领域的重要概念，它可以被定义为一个实体，可以在所处的环境中感知信息，并且根据这些信息作出决策，然后实现特定的目标和任务。智能体有自主性，感知能力和决策能力。其中 LLM 作为智能体的大脑，辅以规划、记忆和工具使用等关键组件。智能体能够将大型任务分解为更小的子目标，并使用短期和长期记忆来处理信息。

智能体平台通常指的是一个允许用户创建自己 Agent 的平台，这样的平台可以提供一系列的工具和服务，帮助用户快速构建和实现AI应用。

AI Agent 被公认为 AIGC 发展的确定性方向。大语言模型目前更加擅长对话和绘画类任务，导致 LLM 能力普遍存在固化。同时由于算力问题，LLM 的记忆里有一定的限制。而 AI Agent 有规划能力，可以将大任务拆解为自任务，并且可以自动化使用和调用工具，为大语言模型的应用带来了更广阔的空间。



▲ 图 12 AI Agent架构示意图

2.3.6 AI资产运营

AI 资产通常指的是在人工智能领域中，由企业或个人所拥有的、能够产生价值的资源或技术。第一步是 AI 资产的管理，确保算力可以有效利用，快速训练出符合业务场景需求的模型的关键。第二步通过 AI 资产的运营，让 AI 资产实现共享，可以轻松下载和训练组织之间的预训练模型，大幅节约研发者的模型训练成本和时间，构造组织内部 AI 的开源社区。这些资产包括但不限于以下几种类型：

1.数据集市：

数据是训练模型的基础。高质量的数据集可以提高 AI 系统的性能和准确性，且大模型数据庞大，开源数据集需要被登记和管理，平台需要实现组织内部的数据共享，为大模型训练提供语料库。该模块具备如下的能力：

a) 数据上云：

· 模块支持网站数据、专业文献、行业数据等多种安全数据源连续地导入数据。

· 支持对各种原始格式的数据格式，例如 PDF，DOCX，XML 接入平台，对结构化和非结构化的数据导入，实现大规模和大体量的数据上云。

b) 数据管理：

· 对数据的来源和权限，以及元数据属性进行管理。

· 对数据的业务属性和 AI 属性进行管理，例如数据集的领域，应用场景和相关权限。

· 对数据集提供上架，更新和下载的能力，对数据集全生命周期管理。

c) 数据加工：

· 平台提供以 Python 和 SQL 为基础的数据加工能力，提供特征工程服务。

d) 数据标注：

· 平台提供集成的标注工具，支持不同类型数据的标注需求，如图像、文本、音频和视频。提供直观的用户界面，使标注人员可以轻松地对数据进行分类和标记。

· 设计和实施标准化的标注流程，确保数据标注的一致性和准确性。且支持多人协作标注，实现标注任务的分配、审核和质量控制。

· 建立标注结果的反馈机制，允许标注人员根据模型训练的反馈调整标注策略。

· 自动化标注，利用机器学习技术，开发自动化标注工具，以减少人工标注的工作量。

e) 数据展示：

· 平台提供基础的BI报表建设能力，允许对数据集的相关结构化信息开展业务分析，对核心的数据指标进行可视化报表展示。

2.模型集市：

模型集市支持用户发布和下载开源的预训练模型。实现对模型共享和快速模型的部署。同时将用户训练好的模型进行上架、更新、版本管理，实现对模型的全生命周期管理。

a) 模型注册：

· 模型注册提供模型的上架能力，对模型的版本进行控制，快速完成模型的业务打标，如来源、应用场景、描述说明等。可以让用户快速找到对应的模型并且完成应用授权。

b) 模型部署：

· 模型部署是快速地将上架的模型部署到GPU计算资源，涉及到模型的容器化和推理权限配置等。对模型在实际应用中的性能和可靠性进行测试。

c) 模型库管理：

· 模型库提供一系列的开源预训练模型，这些模型可以针对图像识别、自然语言处理、推荐系统等进行优化。包含一系列的模型文档，用户反馈系统，调用次数和下载次数的监控。

3.镜像集市：

提供丰富的镜像资源库，允许用户浏览，选择和下载各种大模型训练需要的环境依赖。该模块通过提供预配置的镜像，显著简化了大模型训练和部署的复杂性，降低了工程实施的门槛，支持开发者在高效、可靠和安全的环境下进行大模型的开发和创新。

a) 镜像导入：

· 需要为官方镜像导入和用户镜像导入提供标准和验证流程，确保镜的质量和安全性。同时制定一套镜像命名规范、标准化镜像上架流程。

b) 镜像库管理：

· 建立流程服务和管理镜像库，包含关键镜像的官方源更新和软件更新，同时允许用户镜像访问权限控制和共享用户镜像。

c) 镜像诊断：

· 对于大型和复杂的模型，提供镜像诊断工具，帮助用户排查和解决镜像使用中的问题，并提供核心技术支持。

4.实验集市：

实验集市为研究人员提供了一个平台，用于管理、共享和协作实验流程和结果。基于算子和工具实现不同场景的算法业务流，对实验设计、执行、结果分析和共享。

a) 工具管理：

· 提供一个集成的工具管理平台，允许研究人员访问和管理各种实验工具和软件。同时支持工具的版本控制和依赖管理，确保实验的可重复性。

b) 实验管理：

· 实现实验的全生命周期管理，从实验设计、执行到结果分析，支持实验的自动化执行。

c) 实验工具研发：

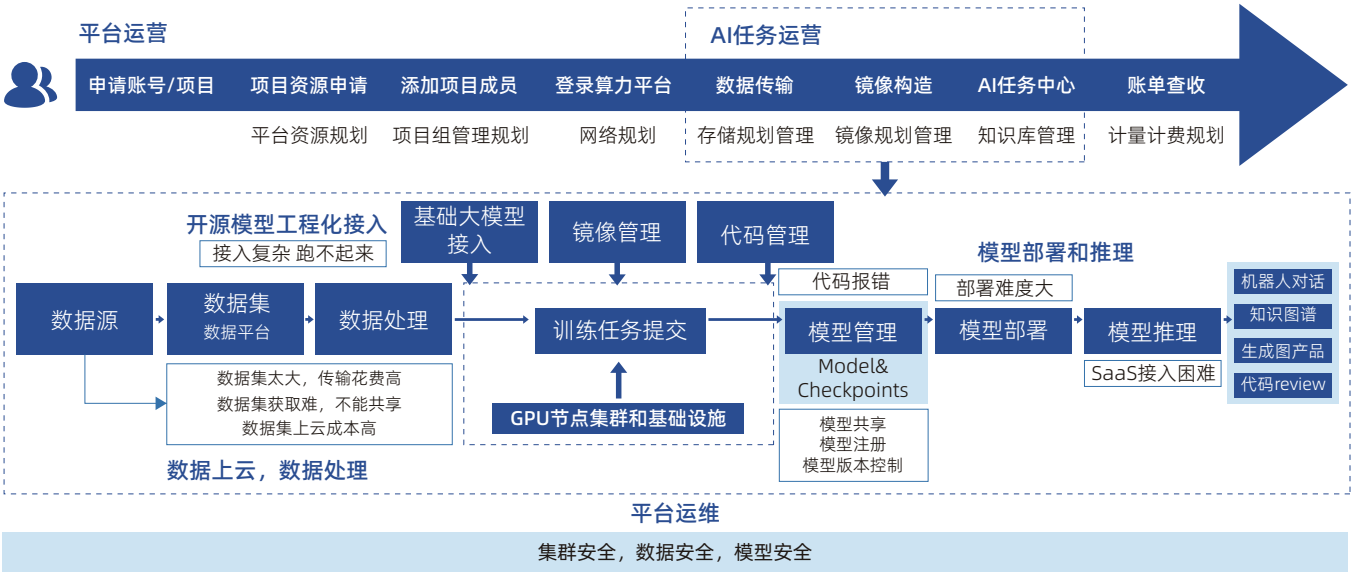
· 鼓励和支持研究人员开发新的实验工具，以满足特定的研究需求。提供工具开发的资源和指导，促进创新和协作。

2.4 智算平台运营

智算平台的运营从用户的使用需求开始，覆盖资源使用的全流程，形成智算平台运维运营体系，由用户运营、资源运营和运营管理三个方面构成。

2.4.1 用户运营

对智算平台运营的范围和内容从沿着用户使用路径展开，具体包括AI常见问题技术支持、知识库、培训以及费用管理和账单查收等方面。



▲ 图13 智算平台用户运营流程示意图

1.用户管理

通过产品化的能力以及相关的运营运维流程，为用户提供一系列针对智算平台使用的服务，让用户可以高效地管理自己的账户、资源和服务。同时运营团队需要制定一系列的规范和机制，指导用户高效地使用算力。其中包含：

1.用户和项目组注册：为用户提供账户的注册、项目组注册和管理等能力。

2.资源开通：为用户提供资源和规格的选择、开通算力和存储资源。

3.订单管理：用户可以管理账户的资源订单，以及上传和编辑合同模版。

4.工单管理：用户可以提交和跟踪工单，以及查看故障待办和当前进展。

5.账户资金管理：用户可以充值账户、查看资金余额、资金使用明细，以及管理账单和发票。

6.消息和通知：用户可以接收和查看系统消息，以及工单状态更新。

7.用户信息和安全：用户可以维护个人信息，如修改密码、绑定手机号和邮箱，保障账户安全。

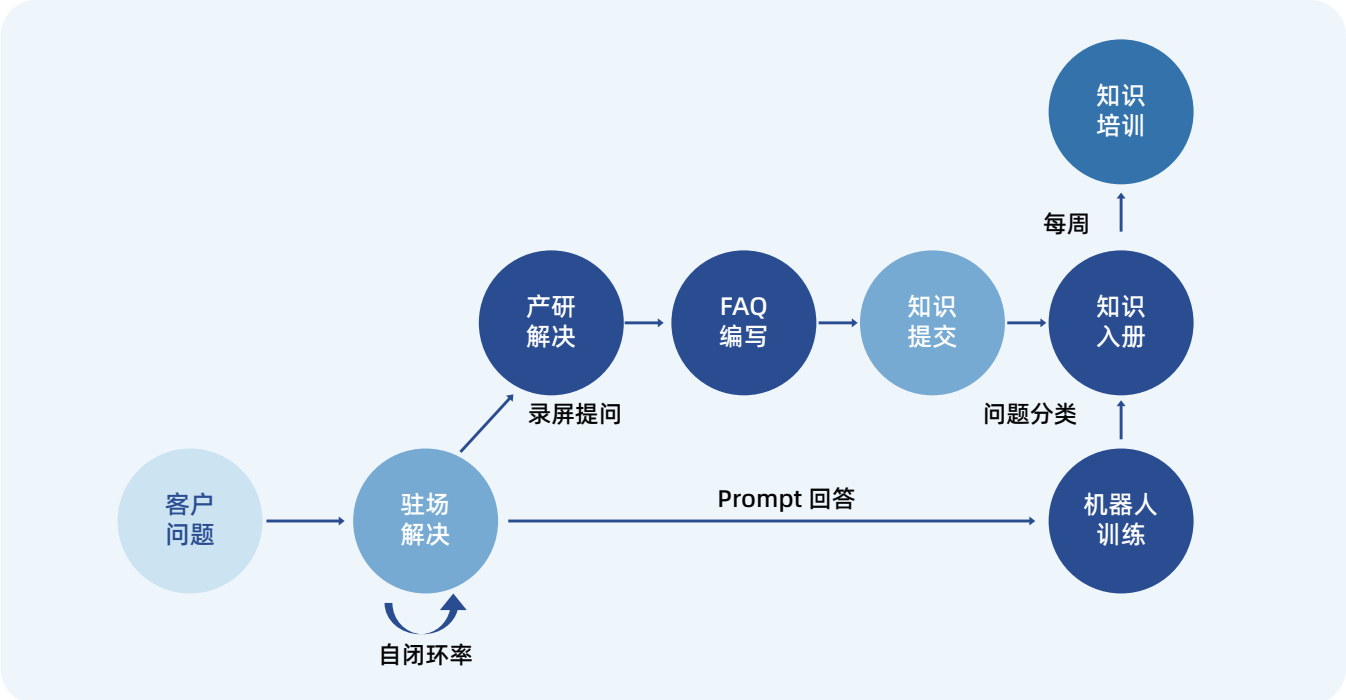
2.工单答疑

智算的工单答疑是运营难度最大的模块，用户的问题遍及平台运营、AI 模型、基础设施和上层应用。理解数据模型的用户可能看不懂网络硬件，传统运维能力难以满足 TensorFlow、PyTorch 等框架的使用问题解答。智算平台运营是云计算、大数据、人工智能平台运营的结合，要求运维人员在回答和理解客户各种的问题的时深刻理解 Linux 底层操作系统、K8s、深度学习以及镜像等专业内容。

工单答疑是影响平台客户满意度的重要服务模块，可采取根据用户画像分群体运营的模式，例如群运营、VIP 服务运营等。工单答疑需要通过对工单分类、对用户意图进行统计，从而对常见的问题及解决方案文档进行总结，通过训练自主问答机器人将结构化的正确答案输出。

3.知识库编写

知识库主包含各个产品的研发手册和使用说明、用户使用手册以及常见问题 FAQ。知识库由专门的知识管理人员做统一编写，确保知识库的质量。知识库编写完成后也需要开展知识培训给对应的用户和工作人员。



▲ 图 14 运维运营 FAQ 知识库生产流程

2.4.2 资源运营

智算运营平台根据不同类型任务对算力资源需求，提供算力纳管能力。通过建设一个高效、稳定、可靠的智算平台，为用户提供高质量的算力服务。资源运营在支持计算需求和提高资源利用率方面发挥着关键作用。

	任务类型	相关资源考虑
单机 单卡任务	适用于小型到中型的计算任务，如传统机器学习，数据加工、数据分析、图像处理等。	只涉及单卡计算资源，管理和调度相对简单，卡资源需求灵活，但容易造成整台机器的碎片化，影响算力的供给。
单机 多卡任务	适用于更高计算能力的任务，例如深度学习模型的复杂训练。	单机多卡任务可以显著提高算力，需要有效的资源规划，考虑卡内通信效率。
多机 多卡任务	适用于超大型计算任务，如超大规模深度学习模型训练、科学计算。	多机多卡任务需要复杂的集群管理和网络通信策略，确保不同节点的有效协同。

▲ 表 2 平台主要运行的任务类型和相关资源考虑

1.资源纳管

算力运营平台实现了对多样化计算资源的全面纳管，包括多种型号的GPU和CPU和定制化的计算资源。用户可以在统一的交互界面中，轻松管理整个计算平台的服务目录，实现资源的整合与优化配置。

算力运营平台支持从算力资源申请、审批、创建、变更到回收的全生命周期管理动作。平台能够精确记录资源的申请和变更记录、资源的项目归属和资源的计费主体，提供根据资源类型、作业目的、提交者身份等不同维度的资源审批能力，实现对资源的全生命周期运营管理。

2.算力调度

算力调度是指在系统中合理分配和利用计算资源的过程，其主要目的是提高整个集群的利用率，保证任务的高效执行。算力调度系统的复杂性主要由两个因素造成：一是业务资源约束因素；二是底层的基础设施、资源隔离能力约束因素。调度器的一项核心任务就是按照某一策略从集群中挑出最合适的物理机，通过机器混合调度提升机器使用效率。智算集群通过容器化的方式屏蔽了物理机之间的配置差异，进一步提升使用体验。

3.资源池化

传统的算力管理通常以物理机为单位，将物理机分配给对应团队，由相关团队内部再进行资源分配，在资源空闲时造成了极大的浪费。智算平台用虚拟化、负载均衡等技术将计算资源（如CPU、GPU、内存等）集中管理，形成一个统一的资源池，可根据资源余量、用户需求进行动态分配，提供更好的可扩展性和灵活性。同时支持队列管理能力，在资源不足的情况下开启计算任务排队模式，在有资源空闲时自动启动新任务，极大提升了资源利用和流转效率。

4.资源治理

算力的资源治理包含自定义治理策略、全链路资源治理、资源效能等方面。

自定义治理策略

根据采集的指标，结合智算应用场景，搭配贴合实际治理场景的治理策略，更精细、更精准的发现可优化的实例，治理的指标如下表所示：

	现象	原因
资源配置问题	GPU 内存/CPU 占满，但 GPU 卡未被占用	机器的内存和 CPU 被其他任务占满，GPU 实例无法启动
	GPU 卡被占用，但是 GPU 利用率为0	实例不需要 GPU 卡，建议申请低配机器
	在多卡训练任务中，长期有卡闲置	GPU 卡数申请过多，建议释放部分 GPU 卡或者更新代码
用户管理问题	空间长时间无人登陆	项目组无活跃用户
	用户欠费	用户资金不足
任务管理问题	AI 任务实例运行时间过长	运行时间久，无人管理

▲ 表 3 治理指标分类

全链路资源治理

全链路资源治理包括对治理项目的持续监控、智能推送治理建议、详细查看治理记录、实时线上反馈以及持续的校验与巡查等关键环节。运营服务团队能够通过这一机制，获得对治理状态的洞察分析，从而确保治理措施的高效和精确执行。此外，团队成员可以通过任务分配、即时在线反馈、定期巡查以及策略调整等手段，不断推动治理规则的持续运作与优化，形成良性的闭环资源治理能力。

资源效能

资源效能主要是对资源使用情况进行监控，为资源优化和管理提供数据基础和依据，并且开展对应的资源分析。和传统的机器监测重点不同，智算平台重点监测显卡性能指标。在 AI 小模型时代，由通信问题造成的性能瓶颈较为少见，而在 TB 级大模型时代，分布式训练及大规模数据可能会导致训练中断、梯度爆炸、算法重跑等问题，造成时间和成本的损失，因此资源效能模块对任务稳定性非常重要。资源效能治理主要包含以下能力：

1) GPU 性能监控：

· 实时监控显卡性能指标，包括GPU使用率、显存使用情况、温度等，以预防过热和故障。

2) 任务管理：

· 盘点当前运行的任务数量，优化任务队列，减少作业等待时间。

3) 存储监控：

· 监控系统内存和存储的使用情况，确保数据读写不会成为限制因素。

4) 网络通信：

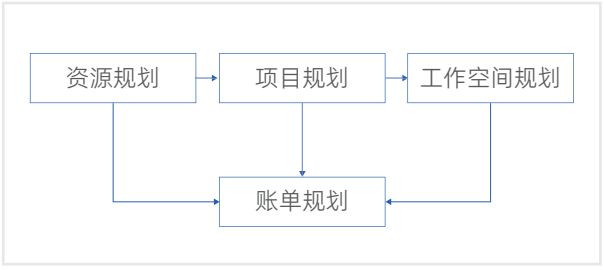
· 分布式训练中的节点间通信是关键，需要监控网络带宽和延迟。

2.4.3 运营管理

运营管理为平台在进行用户管理和资源管理时提供管理者视角，同时也助力平台从 GPU 机器的持有方转为对公众租赁算力的持续利润的 IT 运营中心。运营管理能够更加有效地处理和分析数据信息，给企业组织的决策系统提供信息支持，从而提高对平台整体的运营效率，降低额外的维护和管理成本。运营管理分为运营流程规划、数据驱动的精细化运营以及计量计费三个部分。

1.运营流程规划

运营管理人员需要依赖目前的组织现状和产品形态对计算资源、工作空间、项目组和账单等进行规划，同时也需要为用户使用平台的账号申请、资源申请/升级、工单提交、AI 资产管理、AI 任务提交、缴费、项目失效等提供流程规划和服务支持。



▲ 图 15 运营规划示意图

▲ 图 16 运营流程示意图

2.数据驱动的精细化运营

智算平台需要沉淀用户使用行为与资源运行数据，通过深度分析和挖掘，了解智算平台的运营情况及用户需求，来进行决策和优化，使运营管理团队能够更加精准地了解自身运营状况，及时调整运营策略，提升平台的使用效率。其中可以接入的数据主要包括：

1.平台内用户任务数据，例如用户在智算平台上的日活、任务失败数、工单个数。

2.机器状况，GPU 机器使用率和网络带宽。

3.自动化周报和月报和平台经营分析状况。

4.知识库文章数量，以及知识库浏览和下载量。

5.服务人员能力指标，比如变更次数、缺陷需求数、回答工单个数。

3.计量计费

智算平台计量计费需要提供、账单、对账、发票、代金券、支付、价格管理等模块。

1.订单管理：

支持生成服务订购的订单信息，订单记录服务信息、支付信息、服务开通状态。

2.账单管理：

支持月度账单的统计展示，包含消费汇总金额信息和云产品汇总金额信息。

3.发票管理：

运营方对用户提出的发票申请、撤回等操作进行审核审批以及统计分析。

4.代金券管理：

针对代金券的管理，包含限定适用的产品、用户及订单金额、支持复制代金券信息、控制代金券有效性、快速创建代金券、发放代金券和查看代金券。

5.支付管理：

支付管理支持按不同支付渠道统计查看支付金额的整体数据，包含交易金额、交易笔数、收款和退款的统计。

6.价格管理：

支持基于基础产品价格设置产品目录价格，并且能够按照用户、云产品等维度进行产品售卖折扣设置。

运营团队承担每个月对账、出账、收费核对和处理账单相关的工单等工作。常规流程如下：在智算平台的业务流程中，账单预览允许用户在正式账单生成前查看和确认即将产生的费用。对异常账单进行分析和检查，确保账单的准确性。对账完成后，账单被正式发送给用户，明确其应付金额。用户在平台规定的时间内对账户进行充值。平台进行收费核对，验证用户账户是否有足够的算力余额，如果账户资金不足，且未在补交期限内完成充值，可能会导致账户冻结。用户可在指定时间范围内针对本月的账单提交二次确认申请，运营团队将对申请进行审核和处理，完成整个计费周期的闭环管理。

▲ 图 17 完整计费周期

2.5 智算平台运维

为了追求更高的训练速度和模型性能，大模型训练通常以并行计算的方式进行，会使用数百台 GPU 服务器节点。随着模型规模和训练效率需求的提升，“万卡级”超大集群需求日益增强。在这个计算系统中，每个部件都有概率出现异常，系统越大，整体出现问题的概率越高，例如网络的抖动、板卡的故障、GPU 的故障等不可避免，可以认为服务于大模型的计算集群，稳定性保障是智算平台运维面临的难题。

智算平台运维是一项复杂的系统工作，涉及到硬件的维护、软件的更新、性能监控以及故障排查等多个方面。运维的目标是保障集群的稳定性，以达到事前预知风险、事中快速处理的目标。

2.5.1 计算运维和调度

面向海量数据处理和大规模计算的复杂应用，智算平台可以提供高性能计算任务并行调度框架，需兼容主流的 Kubernetes、Slurm、PBS、LSF 等调度器及多种编程模式，并具备高可扩展性，支持十万以上的并行任务调度能力，支持自动检测故障和系统热点，重试失败任务，保证任务稳定运行。

Kubernetes 先后提供了对不同芯片的集群管理调度的支持，进一步提高了对 GPU 等扩展资源进行统一管理和调度的能力。容器化调度带来如下优势：

1.更加开放：适配开源标准的 Kubernetes 和 NVIDIA Docker 方案。

2.更加简单：优秀的用户体验。AI 应用无需重编译，无需构建新的容器镜像进行 CUDA 库替换。

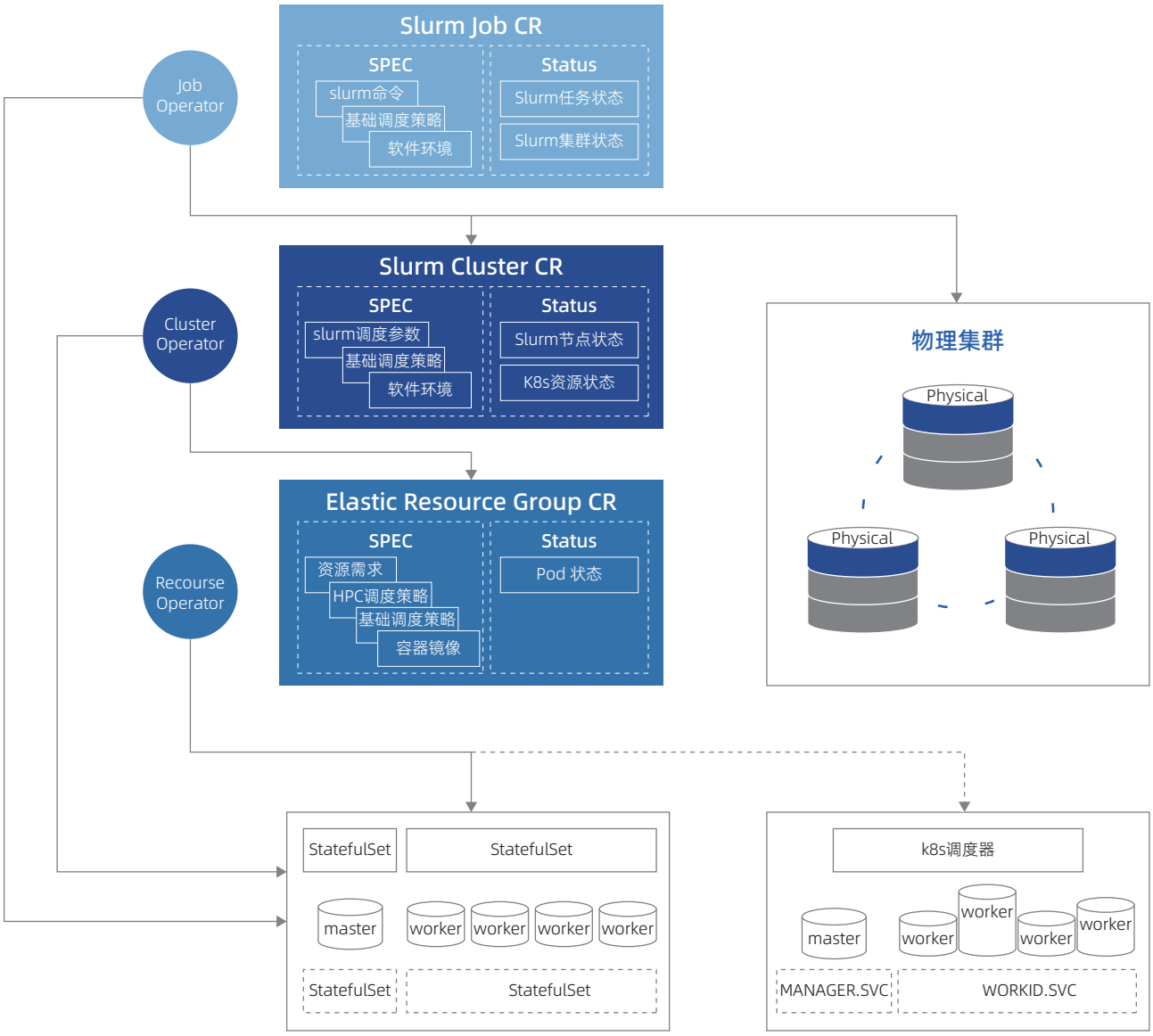
3.更加稳定：针对 NVIDIA 设备的底层操作更加稳定和收敛，而 CUDA 层的 API 变化多端，同时一些 Cudnn 非开放的 API 也不容易捕获。

4.完整隔离：同时支持 GPU 的显存和算力隔离。

优势	说明
支持共享调度和显存隔离	单 Pod 单 GPU 卡共享调度和显存隔离，常用于支持模型推理场景 单 Pod 多 GPU 卡共享调度和显存隔离，常用于支持分布式模型训练代码的开发
支持共享和隔离策略的灵活配置	支持按 GPU 卡的 Binpack 和 Spread 算法分配策略 Binpack：多个 Pod 会优先集中共享使用同一 GPU 卡，适用于需要提升 GPU 卡利用率的场景 Spread：多个 Pod 会尽量分散使用不同 GPU 卡，适用于 GPU 高可用场景。尽量避免将同一个应用的副本放置到同一个 GPU 设备 支持只共享不隔离策略，适配于已有深度学习应用内已自建应用层隔离能力的场景 同时支持多卡共享和显存隔离策略。
GPU 资源全方位监控	同时支持监控独占 GPU 和共享 GPU。

▲ 表 4 融合调度优势说明

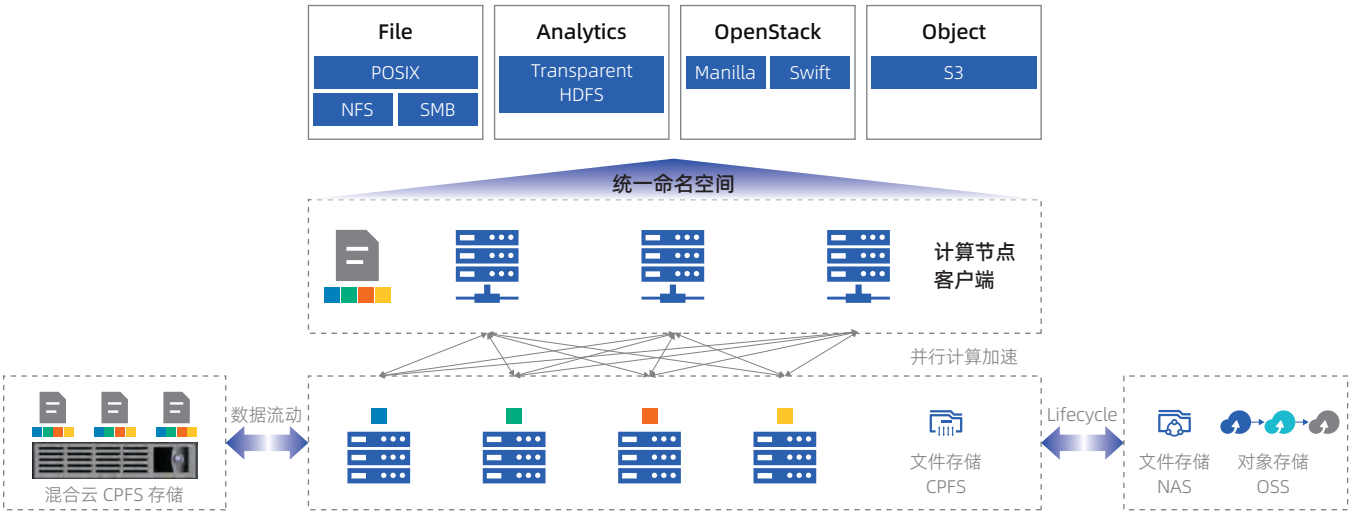
对于 HPC 集群来说，开源 Slurm（Simple Linux Utility for Resource Management) 调度器是主要的选择。HPC 高性能计算集群，主要提供 CPU 算力资源，能够处理复杂和大规模的计算任务，用于解决科学、工程或业务领域中的复杂计算问题。Slurm 具有较高的容错性和高度可扩展的大型和小型 Linux 集群资源管理和作业调度系统。超级计算系统可利用 Slurm 进行资源和作业管理，以避免相互干扰，提高运行效率。



▲ 图 18 slurm资源调度原理示意图

2.5.2 存储运维

智算平台的存储采用专为 AI 计算场景设计的高性能分布式存储。高性能分布式存储可以支持成百上千台机器的同时访问，具有高吞吐、高 IOPS（每秒输入输出操作数）和亚毫秒级的延时。其中，文件系统对于 AI 训练、自动驾驶、基因计算、影视渲染、石油勘探、气象分析、EDA 仿真等场景具有更高的适配性。



▲ 图 19 智算平台存储架构

存储集群在运维过程中常见的问题表现为数据传输慢、离线数据导入存储集群困难、数据迁移难、存储性能监控难等问题。可采用下列运维手段进行处理。

1.数据传输慢：

数据传输慢的问题在实际处理中可能涉及网络、文件类型、硬盘数量等方面：

a) 网络带宽检查：

确保网络连接的带宽足够支持所需的数据传输量，检查是否有其他网络活动占用了大量带宽。

b) 服务器性能：

检查服务器的 CPU 和内存使用情况，确保服务器性能不会成为瓶颈，考虑升级服务器硬件或优化服务器配置。

c) 利用多线程或多进程并行传输数据，提高传输效率。

d) 使用更高效的传输协议：

根据数据类型和传输需求选择合适的传输协议，例如 HTTP/2 或 QUIC。

2.离线数据导入：

离线数据导入大数据集群是一个常见的数据集成任务，分为以下几个步骤：

a) 数据准备：

确保数据格式符合大数据集群的要求，例如 Hadoop 的 HDFS 支持的数据格式。

b) 对数据进行清洗和预处理，以确保数据质量。

c) 数据压缩：

在导入之前，对数据进行压缩，可以减少传输时间和存储空间。

d) 选择合适的导入工具：

根据大数据集群的类型，选择相应的数据导入工具，例如 Hadoop的distcp（分布式复制），Apache Spark 的 DataFrameAPI 等。

e) 网络传输：

使用高速网络连接将数据从源位置传输到大数据集群的节点。

f) 数据导入：

使用大数据集群提供的数据导入工具或API将数据导入到集群中。

g) 数据分区：

根据数据的特性和查询需求，对数据进行合理的分区，以优化查询性能。

3.数据迁移：

存储集群特有的数据流动功能可以实现将对象存储中的数据合并入高性能存储，同时进行统一命名空间的元数据管理。运维策略中应包含数据流动的管理，以确保数据在对象存储和存储集群之间的高效迁移和访问。

4.性能监控：

存储的性能监控可以监控关键指标的状态和历史趋势。指导用户关注数据传输过程中的优化。例如使用GUI开展下面的监控：

优势	说明
Monitoring -> Statistics	提供多种性能图表，展示系统资源以及文件系统的性能。可以选择所需的图表并根据过滤条件监控相关性能指标，还可以在图表上进行平移和缩放，并显示过去的统计信息。
Monitoring -> Dashboards	提供易于阅读的实时用户界面，以图形方式显示关键性能指标的状态和历史趋势。
Nodes	提供一种简便的方法来监控云平台CPFS集群中所有节点的性能、运行状况和配置信息。
Cluster -> Network	提供网络组件的详细性能、运行状况和配置信息。
Files -> File Systems	提供单个文件系统的性能、容量和运行状况方面的详细视图。
Storage -> NSDs	提供单个网络共享磁盘（NSD）的性能、容量和运行状况方面的详细视图。
Storage -> Pools	提供存储池的性能、容量和运行状况方面的详细视图。
Files -> Filesets	提供 Fileset 及其容量信息。

▲ 表 5 GUI页面对应功能说明

2.5.3 网络运维

在智算平台运行过程中，由于涉及到的数据量大、计算密集型任务频繁数据传输和交换，因此需要大带宽、低延迟的网络传输协议。RDMA 网络在保证高速传输的同时，还能减少网络负载，提高数据传输的可靠性，是智算平台和智算中心建设中重要的技术能力。

在执行大规模并行计算任务时，如AI模型训练和科学模拟等，一个高效的GPU集群网络架构一般需要关注以下几个问题。

1.高带宽：

GPU 集群中的节点需要高速数据传输能力，以支持大量数据的快速移动和处理。

2.低延迟

网络通信的延迟需要尽可能低，以减少计算任务的等待时间，提高整体的计算效率。

3.可扩展性：

网络架构应能够随着集群规模的增长而扩展，无论是增加更多的 GPU 节点还是提高单个节点的 GPU 数量。

4.高吞吐量：

网络应能够处理大量并发连接和数据流，保证在高负载下的性能稳定。

5.容错性：

网络设计应包含容错机制，以确保在部分网络故障时，集群仍能继续运行。

6.拥塞控制：

有效的拥塞控制算法可以防止网络过载，确保数据传输的稳定性。

常见的GPU集群网络会将存储网络和计算网络进行分离，完成计算和存储的相互独立运行。

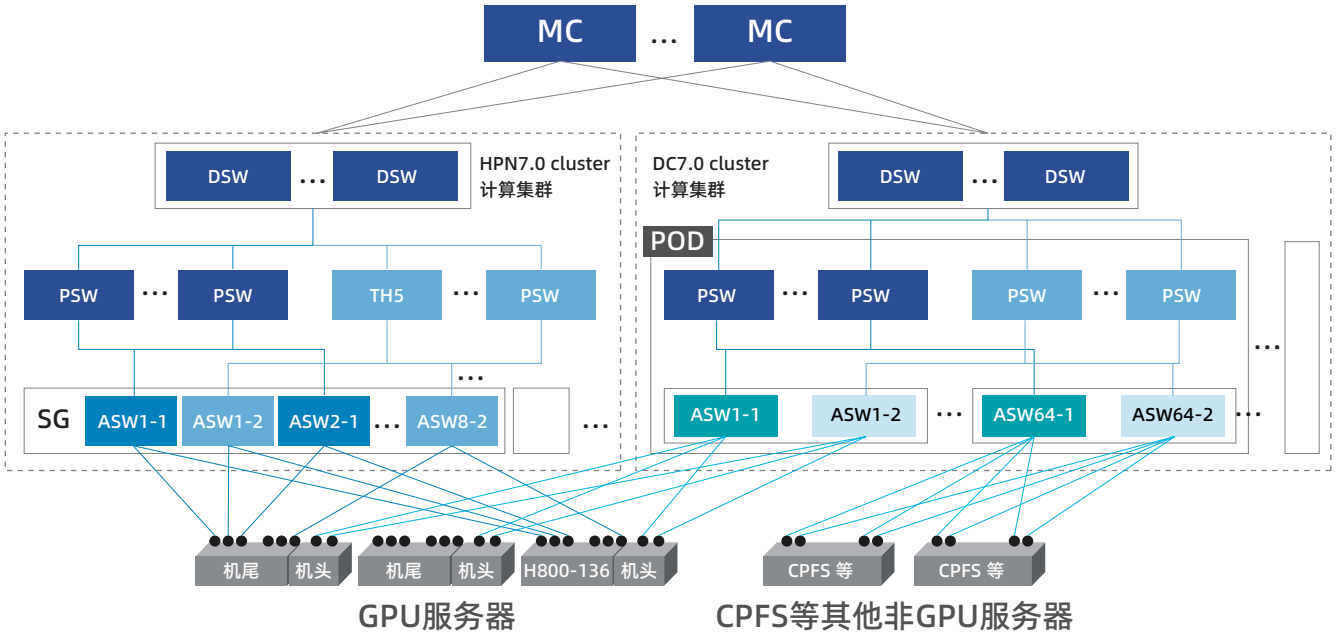
2.5.3.1计算网络

网络集群建设过程中，在保证集群稳定性和性能要求的基础上，往往会在可扩展性和经济性上做一定的取舍。目前主流的网络集群架构分两层 Spine-Leaf 和三层 Core-Spine-Leaf 架构。

以一个千卡 A100 集群共 128 台 GPU 设备为例：计算网络采用 Spine-Leaf 两层架构，32 台 Leaf 交换机加 16 台 Spine 交换机等于48台交换机，每台交换机有 64 口 400G 端口。集群规模从 128 扩大到 256 时，不能简单的做设备的增加，通常有两种处理方案：

一是沿着 128 集群所采用的 Spine-Leaf 两层架构，简单扩大到 256 集群，这种方案的优点是简单、省钱，但两层的 256 集群已经是极限，未来如果要继续扩容会比较麻烦。二是方案是采用 Core-Spine-Leaf 三层架构，前期在网络设备、跳线上的投入相对方案一会更多一些，但为将来扩大到 512 集群提前打好了基础。

GPU 网卡直连到置顶交换机（leaf），leaf 通过 full-mesh 连接到 spine，形成跨主机 GPU 计算网络。如下为 Core-Spine-Leaf 架构 GPU 到接入交换机（leaf）ASW 拓扑联线说明：



▲ 图 20 计算网络架构示意图

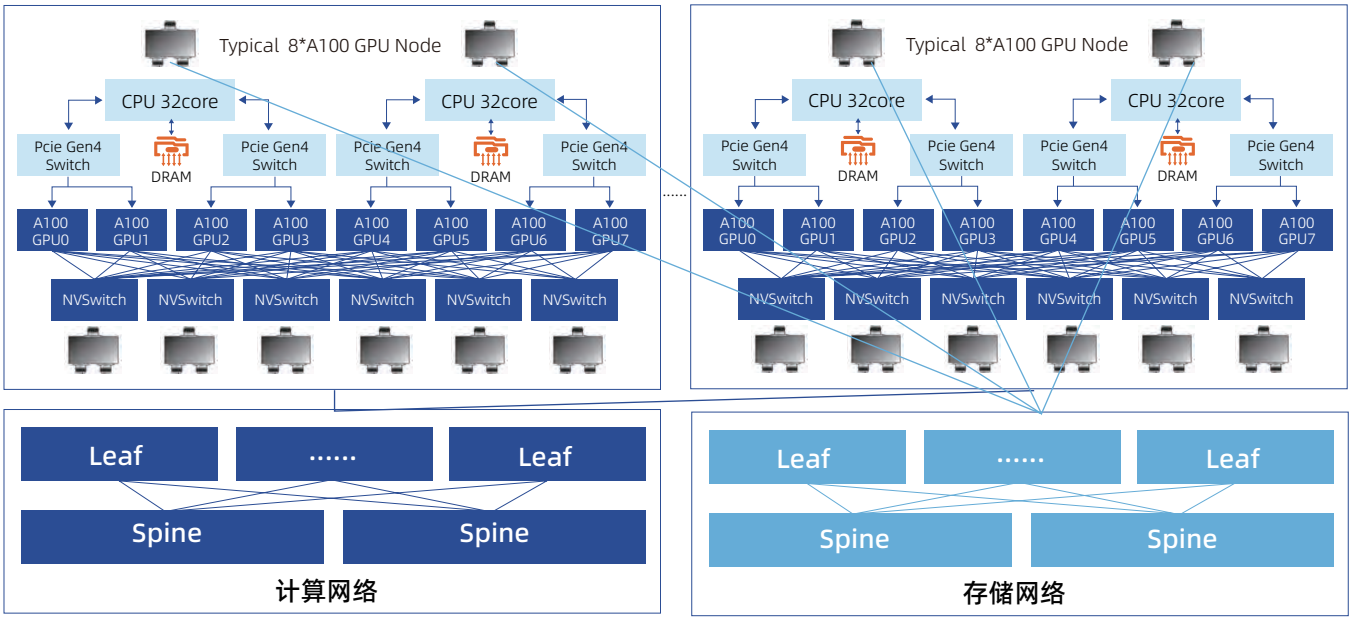
将 256 台 A100 台服务器分成两组 128，每组 128 使用 32 台 Leaf 交换机加 32 台 Spine 交换机，加上 32 台 Core 交换机，一共要用（32+32）*2+32=160 台交换机。这个方案具备跨良好的可扩展性，当集群规模从 256 扩展到 512 的时候，不用重新布放 Spine-Leaf 之间的跳线。

2.5.3.2存储网络

存储网络设计一般通过直连 CPU 的两张网卡，集成连接到一张独立的存储网络环境中，主要的业务目标为：从分布式存储读写数据，例如读训练数据、写 checkpoint 等和正常的 node 管理、SSH、监控采集等。

存储网络性能及功能设计需要考虑采用多协议融合和自动分级存储技术，实现存储空间的高效利用和数据的高效流动，设计时需要考虑网络的高可靠性和安全性，确保不同业务、不同安全级别、不同租户间的数据隔离和互通控制。

为满足大模型训练对于存储高吞吐性能需求，基于全局文件系统技术，可支持超千卡节点扩展规模，为大模型训练提供百 PB 级全闪存储大集群能力，从闪存密度、数据面网络、并行客户端和对等通信机制等多个维度全面提升存储系统性能，智能算力利用率提升 20% 以上。



▲ 图 21 存储网络架构示意图

2.5.3.3数据传输网络

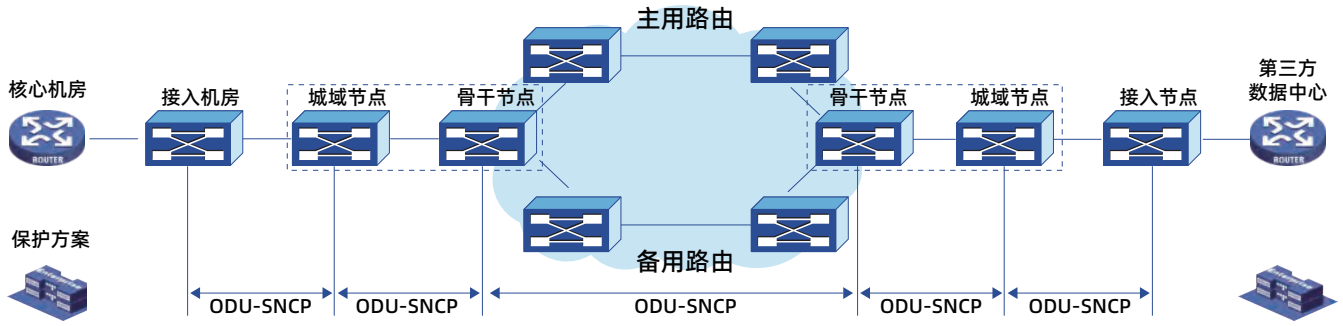
高性能传输网络的设计相对复杂，需要综合考虑多个因素，并根据具体的应用场景和业务需求开展定制化设计，如下是设计数据传输网络过程中需要考虑的内容。

1.需要选择合适的网络拓扑，如星型、环型、网状或胖树（Fat-Tree）拓扑，以满足不同的性能和扩展性需求。

2.确保网络具有足够的带宽容量来处理数据传输需求，避免拥塞和瓶颈。

3.实现 QoS 策略，确保关键数据传输的优先级和带宽分配。

以某智算集群为例，高速数据传输网络在建设过程中提供了1对全程不同物理路由的 100G 高带宽 OTN 线路来组建用户网络。完整的 DNS/DHCP/IPAM 服务，DNS/DH·CP 可支持系统级 HA,实现故障自动切换；建设一套网络应用层流量监控和管理系统，提供 95% 以上 2-7 层协议的识别能力、网络应用性能监测、流量数据展示、IPv6应用可视化等功能；建设一套大数据日志审计系统，提供会话日志记录、流量流向查询、网络性能分析、网络应用性能监测等功能。



▲ 图 22 数据传输网络架构示意图

2.5.4 安全运维

整个安全体系建设的重要参与方包括专业的安全团队，集群产研团队和智算运维运营团队，各个团队充分开展合作，以确保技术基础设施的可靠性和安全性。安全运维负责日常系统维护、软件部署和故障排除。

2.5.4.1智算平台的安全业务特点：

1.资产价值高：

智算平台通常具有丰富的 CPU、GPU 计算资源，对黑灰产业来说是高价值目标。

2.数据敏感：

在算力时代，数据作为商品传输，智算平台中流通着海量数据，涉及医疗、金融、政务以及个人信息等机密数据。若数据遭受篡改或泄露，将造成严重后果。

3.业务形式复杂：

为了便于用户进行使用，智算平台一般会提供便捷多样的数据交互渠道，并且提供可编程的 C/S 模式的 IDE。这种灵活的使用方式会造成暴露面和风险增加并提高安全管控难度。

2.5.4.2基于业务特点的安全需求：

从智算平台的特点出发，开展安全基础设施建设、纵深防御的反入侵体系建设、数据安全建设：

▲ 图 23 智算平台安全能力

2.5.4.3安全建设思路

从反入侵方面的产品层面来说需要构建从网络-端点的纵深防御反入侵体系，防护和检测能力并重。面对复杂的网络安全态势，不能仅寄希望于防住攻击，还应该预先假设被攻破第一道防线后如何开展入侵检测，下面列出了一些参考的产品类别：

1.WAF：提供 WEB 安全防护能力。
2.NDR：提供全面的网络入侵检测、响应能力。
3.EDR：提供全面的端点入侵检测、响应能力。
4.防火墙：提供四层暴露面收敛能力。
5.堡垒机：提供安全运维能力。
6.SIEM：提供整体日志采集、分析能力。
7.容器安全：提供容器安全防护、入侵检测能力。
8.威胁情报：提供基于 IOCs 和 TTPs 的入侵监测能力。
9.蜜罐：提供主动防御、溯源反制能力。

从数据安全的产品层面来说，一般需要产品来支撑数据边界的管控，智算平台的数据安全往往需要结合实际业务开展，实际的业务系统需要具备管控能力和可审计能力，安全产品则居于其次，因此涉及的安全产品不多，下面列举一些参考的产品类别：

1.零信任：提供 VPN 准入、终端沙箱能力。
2.风险和漏洞管理：基于安全产品开展巡检，对发现的风险和漏洞进行治理。
3.入侵管理：预先制定好应急响应流程，基于安全产品开展巡检，对发现的安全事件进行分析、研判、处置。
4.业务上线风险评估：对智算平台的业务、模型、服务开展上线前安全评估，禁止带病带伤上线，带来不可控的风险。

监控告警

平台监控是确保云平台或任何 IT 基础设施稳定、安全和高效运行的关键组成部分。设计一个有效的平台监控系统需要考虑的监控指标，监控指标可以参考核心指标评价模块。

故障处理

故障是指 AI 系统无法正常运行或无法达到预期性能时，导致AI计算平台无法使用或 AI 业务的正常运行受到影响。故障范围包括：AI 基础设施故障、AI 产品故障、AI 业务系统故障，每次故障需要根据影响面进行分级和管理。

级别		影响面描述
P 类平台可用性	P1	关键服务不可用或平台重大异常
	P2	部分平台服务不可用
	P3	产品服务正常，管控不可用
	P4	产品服务正常，管控部分不可用
S 类业务可用性	S1	业务系统对外服务中断
	S2	业务系统功能部分不可用（但服务未中断）
	S3	业务系统受到影响（例如超时、访问慢、重试）
	S4	不影响业务系统运行
I 级基础设施	I1	单机房断电或断网
	I2	大部分服务不可用
	I3	小部分服务不可用
	I4	服务正常，但容量受影响

▲ 表 6 故障等级定义

故障处置标准包含：

1.故障发布时应准备描述故障现象、业务影响、发生时间、报错信息等。

2.故障源自监控、巡检发现以及用户反馈，故障需要确认，对云平台或应用系统的正常使用不造成影响的不判定为故障；计划中变更引起的异常不判定为故障。

3.平台或业务系统发生的任何故障，第一时间通知运维负责人，由运维负责人调度资源进行故障处理，直至故障恢复解决。

4.故障的解决以快速恢复业务为第一优先级，日志的收集、问题分析在事后进行。

5.故障处理过程中根据故障修复情况实时同步处理进度。

故障复盘

故障处理完成后输出故障报告，故障报告应包括故障描述、故障处理过程、故障根因分析、防范措施及整改方案，同时应对整改方案的执行进行监督和检查直至闭环。

重保管理

重保管理旨在客户业务发展关键时间点（如：重大活动/会议、节假日、关键里程碑节点等）对 AI 平台及业务系统提供技术保障，以“重保前排查预防、重保中值守响应、重保后总结复盘”为思想，确保云平台及业务系统的可靠性、稳定性和安全性。

重保范围

1.平台侧：

保障 AI 平台的安全稳定，包括不限于 AI 基础设施、AI 云产品状态、监控告警、应急预案、数据安全及平台风险点等。

2.业务侧：

保障业务系统的安全稳定，包括不限于业务状态、业务压测、业务监控、性能指标、数据安全和备份等。

重保事项

1.沟通/摸底：

通过沟通了解重保背景及诉求，业务侧：业务架构及所涉及产品实例；平台侧：平台状态、产品服务状态，明确重保工作目标。

2.巡检/修复：

重保前要对平台及业务系统进行深度巡检，发现问题并修复，涉及非只读操作严格遵守变更管理规范。

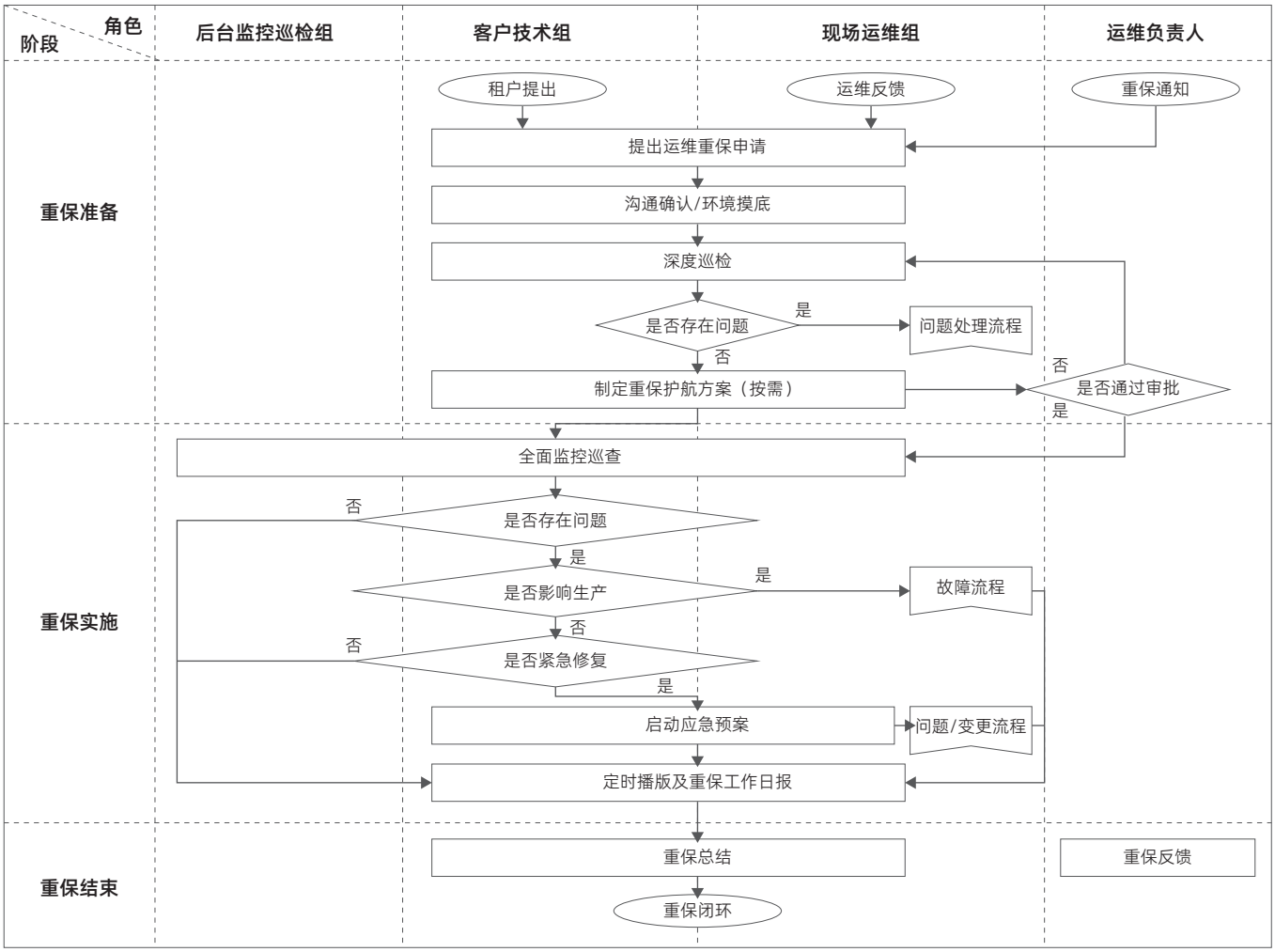
3.制定重保方案：

根据重保事项的重要度按需制定重保方案，重保方案包括不限于业务架构、平台重保巡检方案、遗留问题、风险说明、产品应急预案及执行规范等。

4.权限管控：

重保期间，需要对 VPN 登录环境权限进行管控，只开放重保专用 VPN 账号，其他 VPN 账号权限回收/禁用。

重保流程



▲ 图 24 重保流程示意图

绿色运维

由于 AI 基础设施硬件相对传统硬件需要消耗大量的能源电力资源，为了满足 AI 集群发展的可持续性，机房基础设施的绿色运维显得尤为重要。数据中心 PUE 是目前衡量数据中心用能效率高低的常用指标，无论是国家标准、还是“东数西算”工程，都将 PUE 作为衡量数据中心能效水平的关键指标。PUE 强调的是数据中心用能要尽量用在 IT 能耗上、减少非 IT 类的能耗开销。以全球主要的云计算提供商 AWS 相关的机房能耗指标为例，AWS 的机房遍布全球 26 个地理区域，全球平均 PUE 水平1.12~1.15。

为了满足机房超低的 PUE 设计，AI 集群机房在选址和运维机房基础设施时可以参考如下内容。

1.清洁能源的使用：

机房使用大量使用风电、光伏等清洁能源，这有助于减少碳排放和能源消耗。由于机房所在平均气温较低，数据中心能够利用自然冷源进行制冷，减少了冷却系统的能耗，实现了节能环保。

2.高效的能源使用效率：

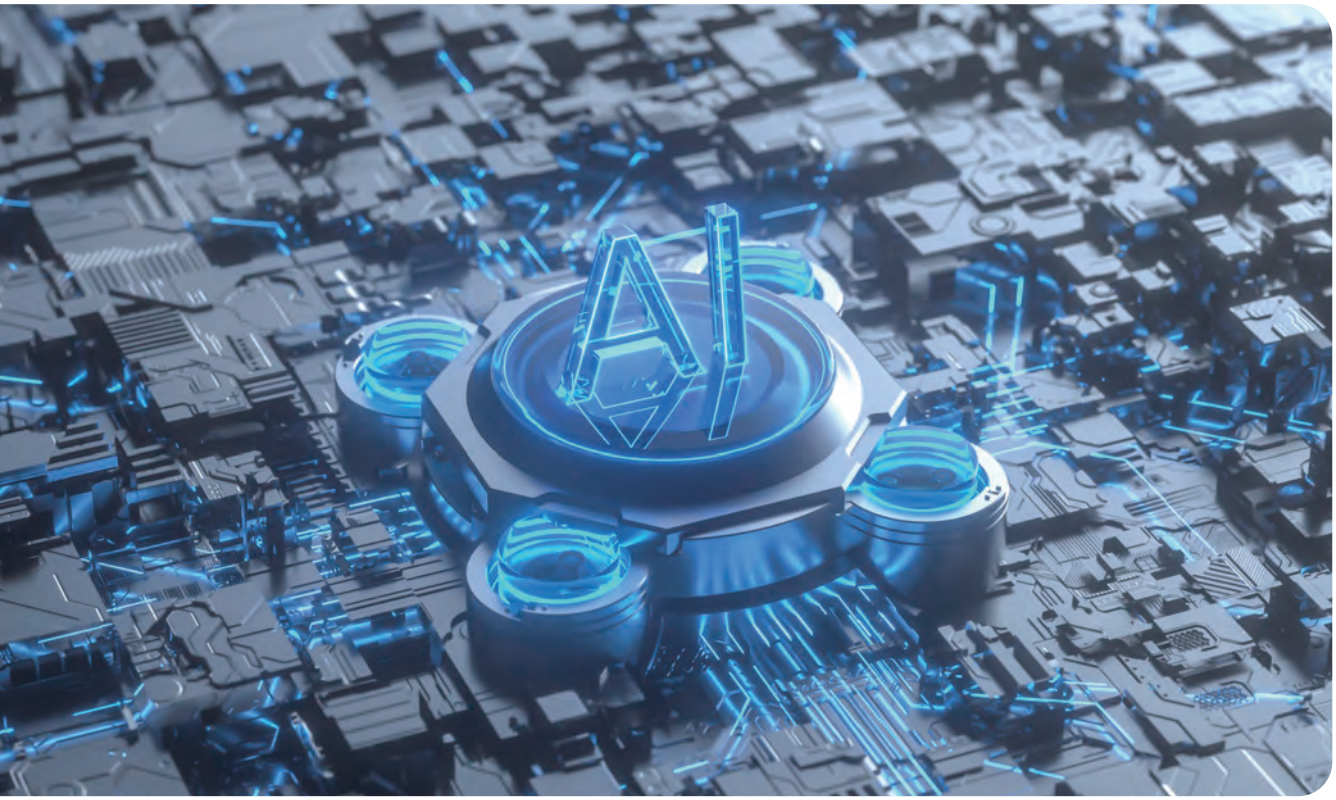
数据中心的 PUE（能源使用效率）年平均可达1.2左右，远低于行业平均水平，表明数据中心在能源使用上非常高效。

3.节能技术的应用：

数据中心广泛使用了液冷、水冷等节能技术，这些技术可以为数据中心节能 70% 以上。

4.智能化管理：

利用人工智能和物联网技术实现智能化管理，提高运维效率，减少能源浪费：机房的地理位置和气候条件为建设绿色机房提供了天然优势，有助于实现低能耗和高效率的算力中心运营。



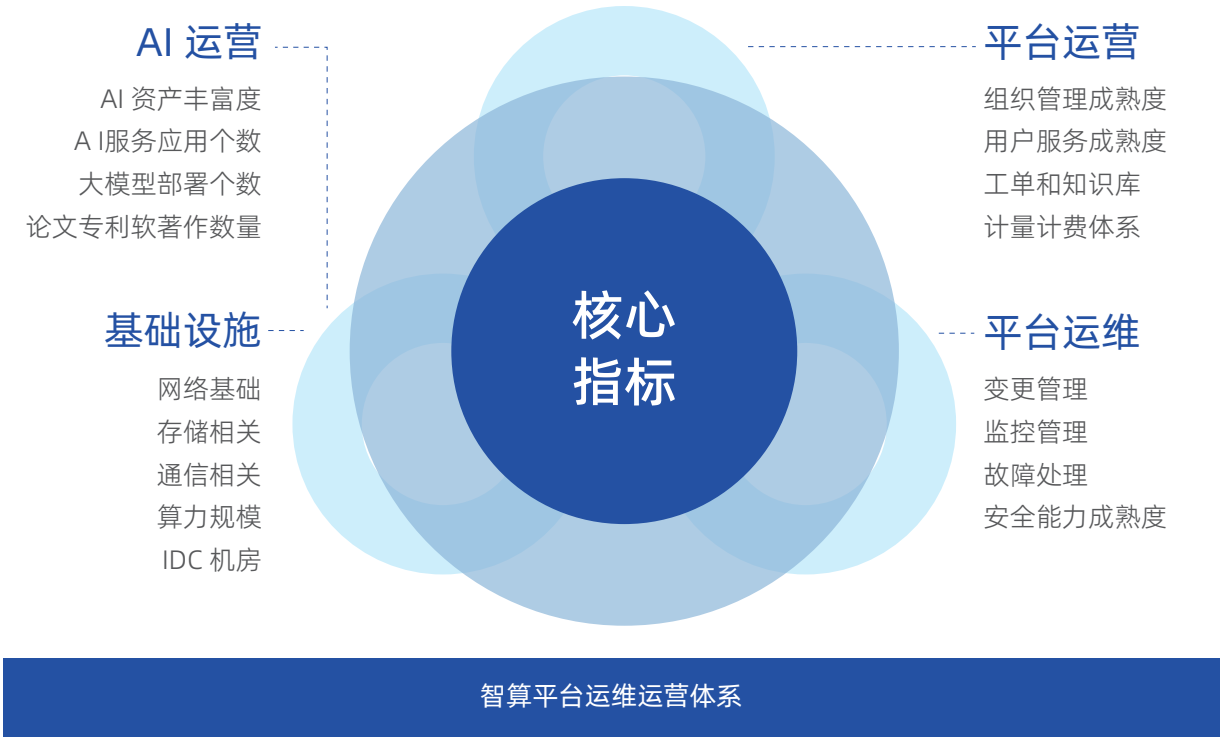
03

智算平台运维运营评价体系及评价指标

智算平台运维运营评价体系及评价指标

智算平台运维运营评价体系的建立旨在提供一个全面、客观的评估方法，来评估智算平台在运维运营领域各细分能力的供应水平。

智算平台运维运营评价体系由四大模块构成：基础设施能力评价、AI运营能力评价、平台运营能力评价和平台运维能力评价，各模块由覆盖软件、硬件、技术、用户体验等指标构成。本评价体系由 4 个一级指标，19 个二级指标，60 个三级指标组成，其中三级指标可以根据实际应用中，数据是否可采集和可评估进行调整。



▲ 图 25 智算平台运维运营核心评价体系

一级分类	二级指标	三级指标	指标解释	单位
基础设施指标	网络	基础网络带宽	平台网络的总吞吐带宽能力	Gbps
		接入交换机 网络收敛比	一组接入交换机下行服务器 总带宽与上行带宽比例	X:Y
		单机服务器 带宽处理能力	单台 GPU 服务器吞吐能力	Gbps
		网络服务响应时间	从用户发起请求到收到响应的时间	毫秒
		数据跨区传输速率	数据在发送端和接收端之间实际传输的速度 通常低于理论最大带宽	Gbps
	存储	存储整体容量	平台存储系统的总容量	以TB或PB计算
		存储吞吐 IOPS	每秒进行操作存储的次数 衡量存储性能的关键指标	次数/秒
		高性能存储占比	高性能存储（如SSD）在总存储中的比例	%
	通信	节点连接外部 网络的带宽	节点连接到外部网络的带宽大小	Gbps
		RDMA 网络吞吐	远程直接内存访问（RDMA）网络的 发送和接收数据量	GB/s
		延迟	数据从发送端到接收所需要的时间	毫秒
	算力	智算能力总规模	平台的计算能力总规模 （在 FP16 精度下来衡量）	FLOPS
		GPU 内存	GPU 节点的内存容量	GB
		CPU 核数	GPU 节点配置的 CPU 核数	C
		服务器个数	平台运行的服务器数量	台
	IDC 机房	正常运行时间	衡量机房可以正常运行的时间比例	%
		电源供应可靠性	机房电源供应的可靠性 可以通过电源故障率来衡量	%
		PUE-绿色	机房能耗指标=机房能耗/用于计算的能耗	
AI运营	AI资产运营	数据集丰富度	包括数据来源的多样性、数据集的大小、 数据质量（准确性、完整性）和数据下载次数	个数/TB/%/次
		模型丰富度	接入的开源模型个数、模型迭代速度、大模型 最大参数量、模型下载次数和模型可维护性	个数
		镜像丰富度	镜像数量、镜像更新频率和镜像下载次数	个数/（每周/月）/ 次数
		工具丰富度	工具个数、工具更新频率和工具下载次数	个数/（每周/月）/ 次数
	AI服务运营	算力应用个数	平台提供的算力应用数量 包括模型应用个数和更新频率	个数/（每周/月）/ 次数
		模型部署	服务响应时间、平均模型部署时长、 最大模型服务并发数和模型调用次数	QPS/时长/ 个数/次数
		微调服务效率	模型微调效果（性能提升）和 模型微调所需时间	%/时长

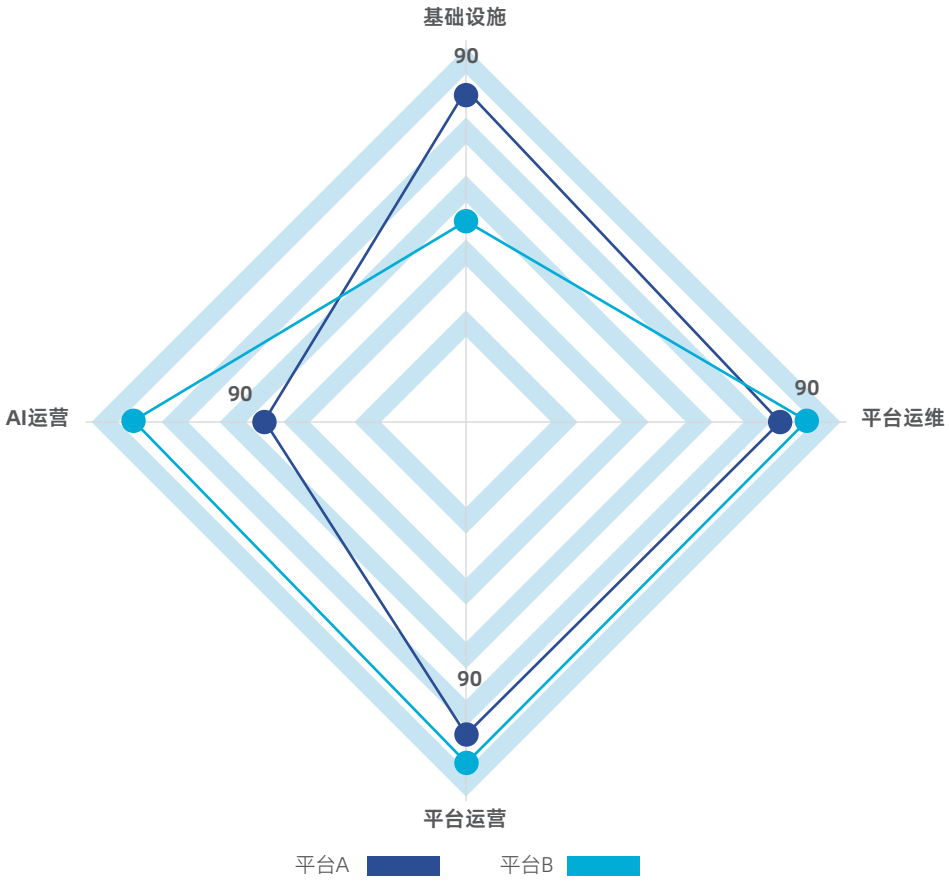
一级分类	二级指标	三级指标	指标解释	单位
AI运营	荣誉相关	专利数量	使用智算平台产出的获得的专利数量	个数
		论文数量	使用智算平台产出的论文数量（A类）	个数
		软著数量	使用智算平台产出的软件著作的数量（A类）	个数
平台运营	组织管理	服务团队人数	智算运营运维团队服务人数	人数
		服务能力专业度	服务人员的工作年限，资格证书， 服务过的项目个数等	个数
		运营运维流程 管理成熟度	团队的管理效率和流程优化 通过流程化个数来衡量	个数
	用户服务	行业覆盖	平台服务覆盖的行业数量	个数
		平台UV	平台的独立访客数量	
		用户满意度	用户对平台服务的满意程度 可以通过调查问卷来衡量	分数
		用户投诉数量	用户对平台服务的投诉总数	个数
	工单处理	工单数量	提交的工单总数	个数
		工单解决率	成功解决的工单占总工单的比例	%
		工单解决时长	解决工单所需的平均时间	时长
	知识库	知识库数量	知识库中的文章或条目总数	个数
		知识库更新频率	知识库更新的频率	每天/周/月
		知识库浏览数	知识库被浏览的次数	次数
	计费	计费功能完整性	计费功能是否自动化 功能完整性	%
		账单运营人员个数	账单运营人员的个数	个数
	培训	累计培训次数	进行的培训的场次总数	次数
		培训覆盖人数	参与培训的人数	人数
平台运维	变更管理	变更成功率	成功实施的变更占总变更的比例	%
		每月变更次数	进行的变更次数（每月/周）	次数
	资源管理	算力资源利用率	包括 GPU 使用率、CPU 使用率	%
		算力性能指标	最高算力 MFU（Model FLOPs Utilization） 实际算力计算利用率	%
		存储资源使用率	存储资源消耗	%

一级分类	二级指标	三级指标	指标解释	单位
平台运维	资源管理	算力资源成本优化金额	通过优化节省的算力资源成本	金额
		算力全生命周期运营成熟度	算力资源管理，资源治理，资源运营的成熟度	分数
平台运维	监控	实时监控指标个数	实时监控的指标数量	个数
		告警触发个数	触发告警的次数	个数
		故障个数	发生的故障总数和频率	个数/频率
		告警有效率	有效告警/总告警数	个数
	故障处理	P1P2级别故障占比	重大故障发生的次数和频率	%
		故障响应和解决时长	故障响应和解决所需的时间	时长
		SLA 达成率	服务等级协议(Service Level Agreement, SLA)达成率，是衡量服务提供者是否按照事先约定的服务标准向客户提供服务的指标	%
	安全	系统安全事件个数	触发的系统安全事件总数（每月/周）	个数
		安全漏洞发生次数	发现的安全漏洞个数（每月/周）	次数
		数据安全和加密	数据的安全性和加密措施是否完善从技术、管理、法规等方面进行定性评估	分数
		防火墙、入侵防御系统等安全设备的能力	安全设备的是否具备以及他们的抵挡的攻防次数	次数

注：同时得到的数据统计指标需要进行归一化处理，按照0-100的分数标准化。

▲ 表 7 核心评价指标详细说明

针对评价体系内的一、二、三级指标，通过基于专家评估的层次分析（AHP）方法，得到评价指标体系中每一个一级、二级、三级指标的相对权重。根据实际应用情况，对指标进行权重设置，形成评价结果，分布展现智算平台在基础设施、AI 运营、平台运营和平台运维四个维度的能力。



▲ 图 26 智算平台综合指数示例

根据评价结果对不同维度的数据开展分析可以对智算平台运维运营能力进行定向优化，如平台 AI 运营方面表现较差，但是基础设施和平台运营等方面表现良好，说明智算平台可能存在推广程度不够、存在资源浪费等问题，需要通过市场推广、运营活动、技术改进等方式进行优化。

04

智算平台运维运营案例

AI运营

智算平台运营

智算平台运维

智算平台运维运营案例

本章结合市场上智算平台运维运营经验对 AI 运营，平台运营和平台运维三个方面进行解析。

4.1 AI 运营

4.1.1 案例1:复旦大学的 AI for Science 运营

复旦大学在智算和大模型领域围绕着 AI for Science 开展了一系列的运营，如伏羲天气大模型和世界科学智能大赛等，鼓励研究者通过智算平台挖掘新的科研场景。

1.世界科学智能大赛

复旦大学、上海科学智能研究院联合阿里云等举办的首届世界科学智能大赛，集结来自全球18个国家与地区的一万余名选手，多数团队具有人工智能和基础学科的交叉背景。选手使用复旦大学智算科研平台，在生命科学、大气科学、材料科学、量子化学、流体力学五大科学赛道进行比赛，其中多支团队的得分超出传统方法的结果。

18+
国家地区

11000+
参赛选手

530
支团队分数超过传统方法

500+
知名高校/机构

举办首届世界科学智能大赛

五大赛道

01 生命科学赛道
生物学年的评价与老年病风险预用

02 量子化学赛道
量子化学分子属性预测

03 流体力学真道
基于NS方程的流动同求解

04 材料科学赛道
金属有机框架材料的预测会成

05 大气科学真道
华东区域AI中期天气型

在开放CFFF平台智能算力的基础上，安全共享了多个特色科学数据集，这些数据集包括：

· 如皋衰老与长寿队列最新数据
· 最大规模MOF合成数据集
· 1000万个构象、能量、力等高精度的第一性原理数据

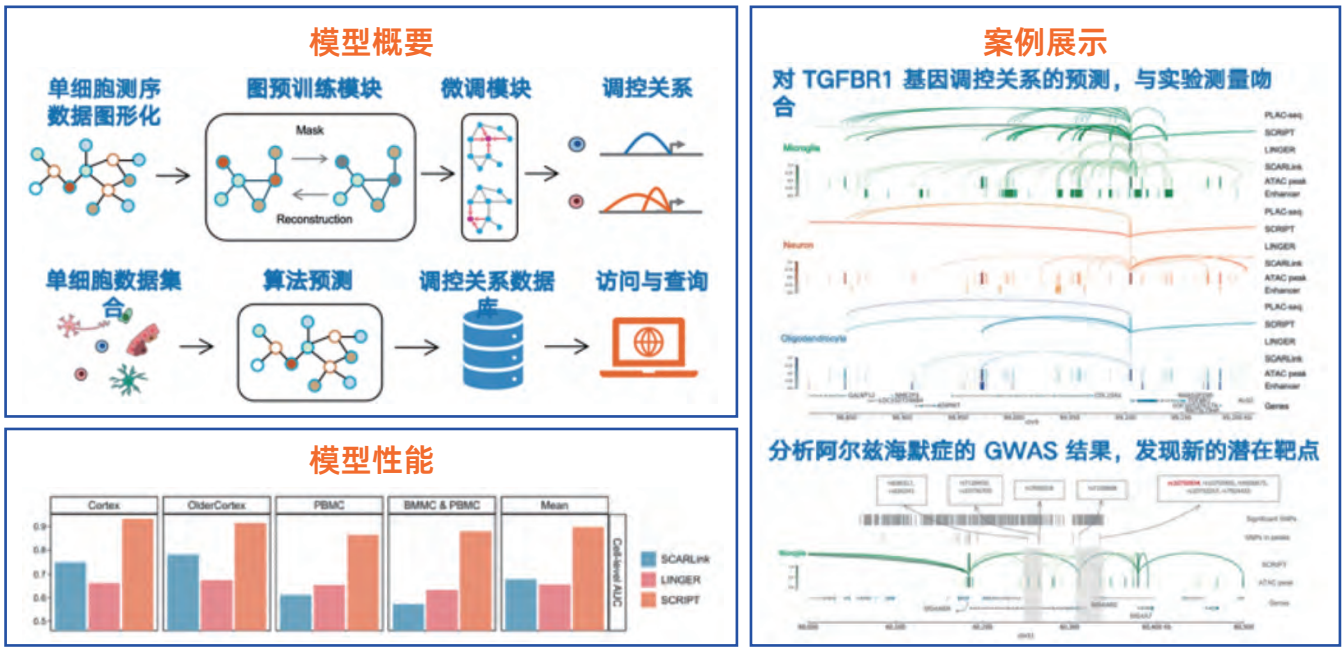
CFFF智算平台
(Computing for the Future at Fudan)

· 全国高校第一的智算集群
· 拥有千卡GPU，8万核CPU，总体算力规模达到40PFlop/s(FP32)

▲ 图 27 智算应用与世界科学智能大赛

2.女娲基因调控大模型

复旦大学联合上海科学智能研究院发布女娲基因调控大模型。该模型基于多组织、多细胞的基因调控关系网络与知识图谱，构建大规模图神经网络预训练模型，首次将基因调控关系以图神经网络的形式建模，实现了单细胞维度调控关系的高精度预测。尤其在调控距离超过100kb的远端调控关系预测任务上，预测精度显著高于发表在Nature Biotechnology和Nature Genetics上的方法，远端调控关系预测精度提升1倍以上，甚至可以预测调控距离超过7Mb的调控关系。目前在阿兹海默症疾病的数据分析中，发现了新的潜在靶点。



▲ 图 28 女娲基因调控大模型及应用

4.1.2 案例2:阿里云 AI 运营实践

阿里云通过端到端的大模型构建服务，提供了一套完整的大模型解决方案，对 AI 资产进行了全方面的管理，同时从应用层到基础设施层，全面支持大模型的研发和应用。

应用层展示了各种 AI 应用案例，如图像识别、文本生成和语言翻译等，有效的发挥大模型在实际应用中的强大能力。

模型层包括 PAI-Easy 系列模型、达摩院大模型系列、达摩院 SOTA 模型库以及社区第三方训练结果。这些模型覆盖了 NLP（自然语言处理）、CV（计算机视觉）和 Speech（语音处理）等多个领域，为用户提供了多样化的选择，满足不同的应用需求，同时打通了 Modelscope，HuggingFace 等 MaaS 平台，支持丰富的模型，为用户提供更强大的数据智能服务。

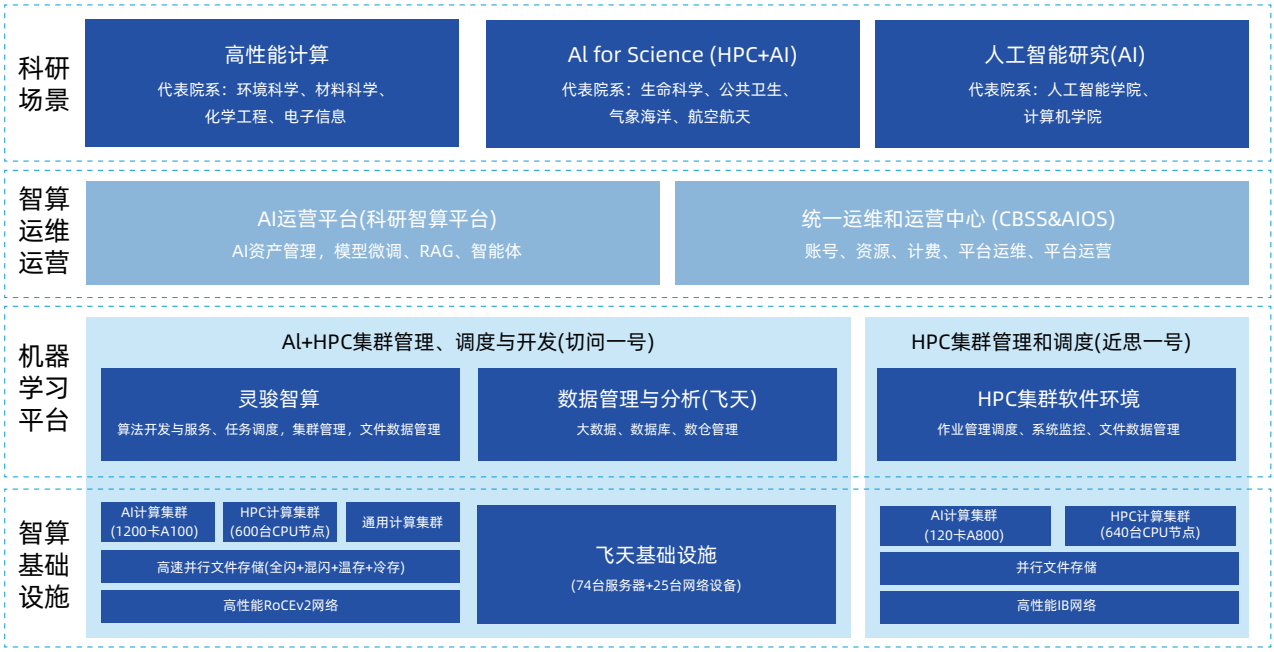
在模型服务（MaaS）平台如 PAI DSW-Gallery、ModelScope 和 HuggingFace，为用户提供了丰富的模型服务和工具，支持用户快速上手并高效进行模型开发。工作层整合了从智能标注、可视化建模、交互式建模、深度学习调度服务、在线模型服务到大模型库服务的多种工具，支持全面的 AI 开发工作流程。通过这些工具，用户可以快速构建和优化模型，大大降低了大模型学习的门槛和开发的压力。

▲ 图29

4.2 智算平台运营

4.2.1 案例1:复旦 CFFF 平台运营最佳实践

复旦大学 CFFF（Computing for the Future at Fudan）平台是我国高校规模较大的云上科研智算平台，由复旦大学联合阿里云、中国电信共同打造，推动交叉学科发展，助力创新路径探索，实现AI前沿技术发展及产业创新。CFFF 平台由面向多学科融合创新的 AI for Science 智能计算集群“切问”一号和面向高精尖研究的专用高性能计算集群“近思”一号两部分组成，算力规模达到 40PFLOPS（FP32），各个级别存储（全闪+混闪+温存+冷存）容量达 70PB。CFFF 平台包含智算基础设施、机器学习平台、和上层的AI运营平台和统一运维运营门户。



▲ 图 30 复旦大学智算平台架构示意图

4.2.1.1 运营工具研发

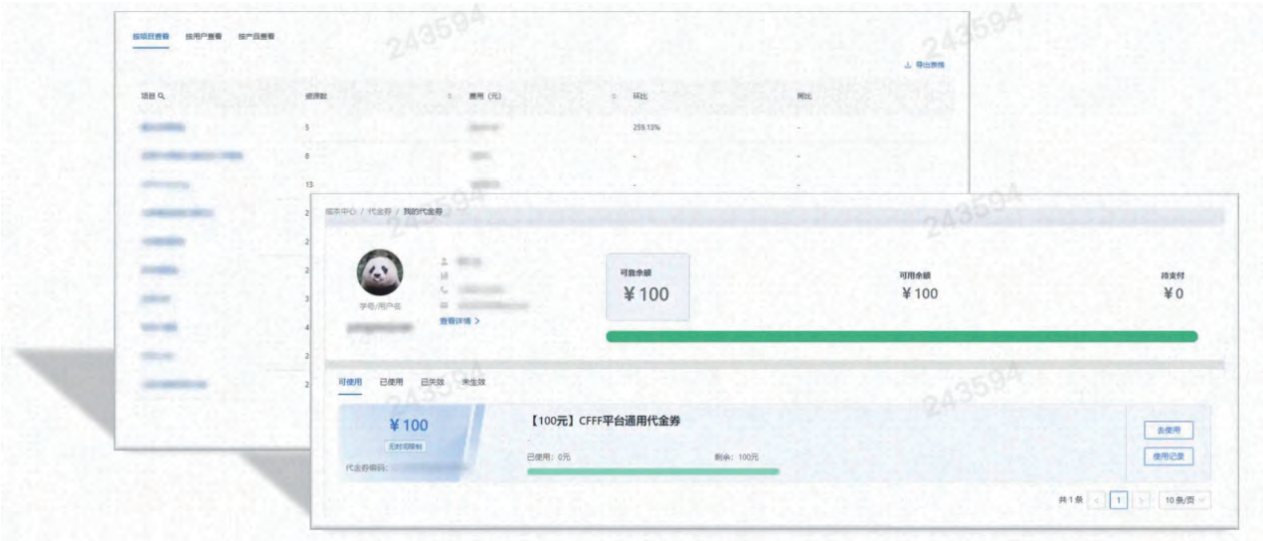
CFFF 平台运营工具可为用户和运营提供产品化的功能，包括机器学习工具、科研智算工具、统一运营工具。机器学习工具提供从数据标注、模型开发、模型训练、模型优化到模型部署的一站式AI开发，支持 PyTorch、TensorFlow 等多种深度学习框架，提供可视化建模、Notebook 交互式建模、跨节点分布式训练等多种功能。科研智算工具提供各学科领域的科研算子，包括生命科学、物质科学、地球科学、流体力学等，通过“拖拉拽”的图形界面，科研人员可以快速的构建实验流程，降低了研究人员使用人工智能技术进行科学研究的门槛。例如，科研人员可利用科研智算工具进行蛋白质结构的预测等。同时平台鼓励科研工具和实验流程的开放共享，以及数据和镜像的共享，通过建设科研开放社区，人员可以访问到更多的资源和数据，加速科学发现的过程。统一运营工具为用户提供平台资源的账户和项目申请、资源管理、账单管理、效能监控、工单管理、运维审计等核心功能。让用户以统一全面的视角来管理各类关键业务流程，提升平台运营效率，降低运营成本。

4.2.1.2 资源运营

资源运营团队负责从资源申请、审批、创建、变更到回收的全链路生命周期管理。平台能够精确记录资源的申请或变更记录、资源的项目归属和资源的计费主体，具备根据资源类型、作业目的、提交者身份等不同维度的资源审批能力，实现对资源的全生命周期运营管理。资源运营团队根据不同任务的性质和优先级，给任务分配合适的计算资源，通过智能调度算法，根据任务的实际需求和当前资源状态，动态地调整任务的执行顺序和资源分配，确保高优先级任务优先执行，并最大化资源利用率。针对算力资源利用的效率低，资源等待时间长的问题需要开展算力的资源治理。算力的资源治理主要包含自定义治理策略、巡检治理策略、市场化收费策略等。

巡检治理策略是指对于重点监控的资源（GPU利用率、存储利用率等），进行定期巡检和治理。GPU 资源通常是按照任务类型进行治理，将资源消耗量大的模型跑在单独的资源池中，同时将单卡训练的小任务，或者还在代码调试阶段的任务跑在另外的资源池中，从而避免资源的碎片化。存储资源的主要治理方向为文件的大小和文件的个数。文件大小主要是较大的文件会占用整体集群的资源，文件的个数(Inode)会限制高性能文件的吞吐。

从复旦大学的运营实践看，市场化收费策略是一种有效的治理策略，对各类资源进行定价和收费可以避免资源的无效占用，提升集群的整体利用率。付费方式分为后付费和预付费两种方式，同时对使用量较大的用户给予一定的折扣。CFFF 平台提供了论文奖励政策，根据用户使用 CFFF 平台产出的论文期刊级别，奖励不同额度算力代金券。

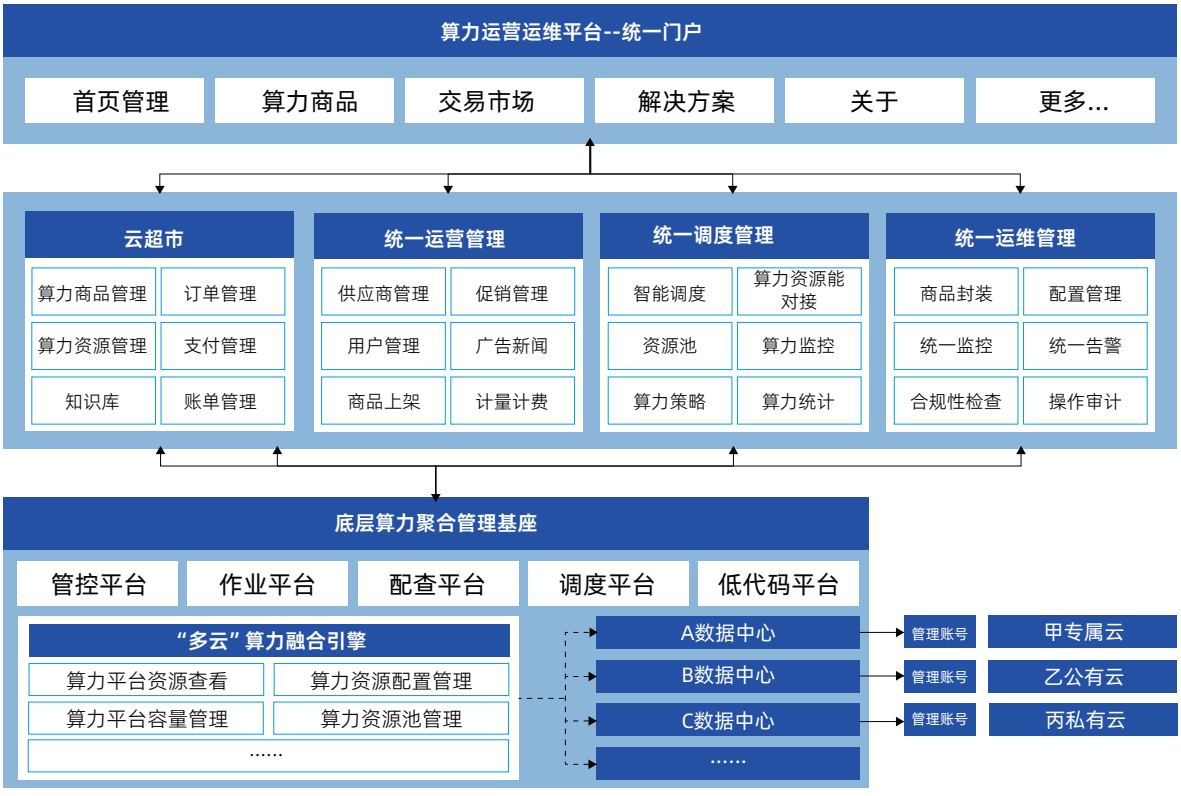


▲ 图 31 CFFF平台代金券模块示例

4.2.2 案例2：骞云算力运营平台

骞云科技是一家 CMP 云管理平台提供商，提供多云管理、混合云管理、云化 IT 资源自动部署及交付等服务，同时也提出了一系列算力运维运营的解决方案，旨在建立先进的算力服务平台，以支持企业和行业在大模型时代的数字化转型和AI升级，该平台基于一系列先进的设计理念，如算力编排、平台工程化思想、全生命周期管理、异构算力池化、一切皆服务（XaaS）等思想建设。

骞云科技不仅提供了基础的算力资源运维和管理的能力，同时为平台管理员提供了服务编排、流程编排和工单设计的能力。例如：算力资源的启停、监控、变更（调整规格、增加磁盘、快照等）、回收等。通过一体化平台的建设实现对算力资源的监控，进而提供集成统一的分析、查询、报告和展示，实现算力资源的统一监控管理、自动化运维管理、全生命周期的资产管理、安全合规、流程管理等功能，帮助算力中心运维管理团队快速、准确、便捷的定位问题，直观快速地诊断和分析问题，将运维模式从被动支持提升到主动管理。同时运维平台抽象了算力资源商品统一建模，进而提升算力资源的动态管理和规范，便于算力供应商或者使用者能够通过统一使用关系来更好地利用算力，提升算力的综合使用率。



▲ 图 32 骞云算力运营运维平台

4.3 智算平台运维

4.3.1 案例1：DataDog 大模型可观测运维

Datadog 宣布推出 LLM 可观测性功能，使得 AI 应用程序开发者和机器学习（ML）工程师能够高效地监控、改进和保护大型语言模型（LLM）应用程序。DataDog 公司与英伟达 GPU 运维的合作形式切入 AI 赛道，负责可视化监控后台 GPU 温度、实时功耗以及工作负载情况。通过算力可视化，帮助企业高效管理、优化算力资源，降低算力成本。

排查大语言模型的程序的问题是一项耗时且资源投入较大的任务。因此，管理大模型任务需要持续监控，以保证持续的性能和安全。其中 Datadog 研发了大语言模型监控工具来解决幻觉，性能和成本，提示词黑客攻击和安全和数据隐私的问题。

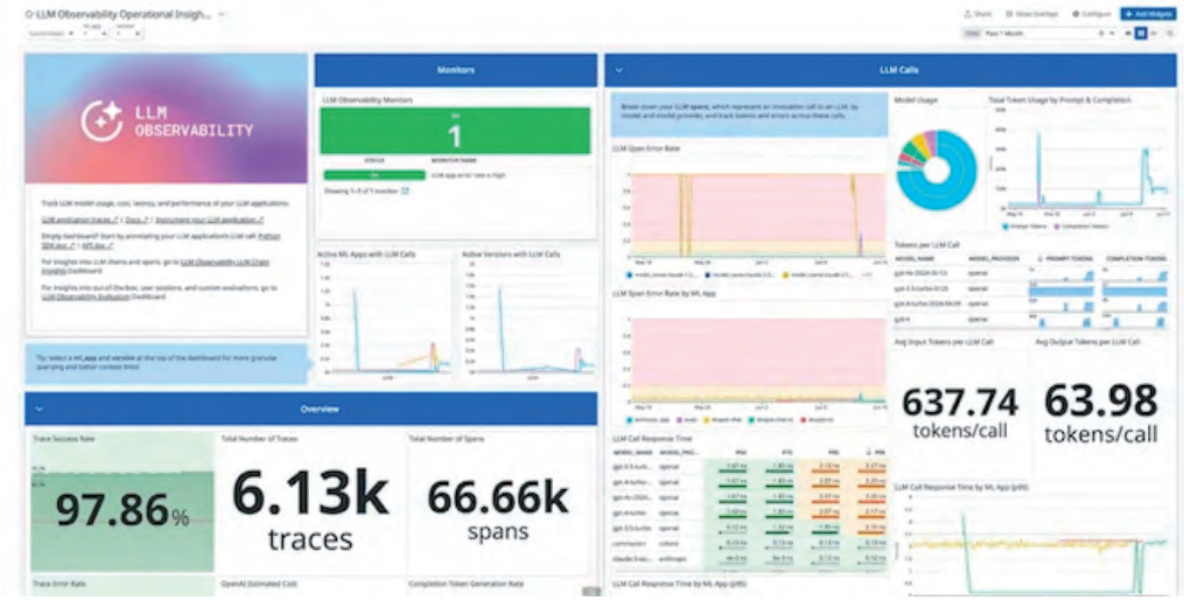
1.大语言模型的推理表现型提高：
LLM 可观测性能够实时监控各种性能评估指标，如延迟和吞吐量，以及响应的质量。通过持续监控这些指标，数据科学家和工程师可以快速识别 LLM 性能的任何偏差或下降。这种主动方法允许及时干预，从而提高模型性能和用户体验。

2.更好的可解释性：
通过 LLM 可观测性，可以深入了解 LLM 应用程序的内部工作。通过可视化请求-响应、词嵌入或提示链序列，LLM 可观测性增强了响应的可解释性。这种增加的透明度使利益相关者能够信任 LLM 应用程序的决策，并识别应用程序输出和逻辑中的任何质量问题或错误。

3.更快的问题诊断：
快速确定 LLM 链中错误和故障的根本原因，全面了解每个用户请求的端到端跟踪。

4.增强的安全性：
LLM 可观测性在通过监控模型行为来增强 LLM 应用程序的安全性方面发挥着关键作用，以寻找潜在的安全漏洞或恶意攻击。通过跟踪访问模式、输入数据和模型输出，LLM 可观测性工具检测可能表明数据泄露或对抗性攻击的异常情况。这种持续监控帮助数据科学家和安全团队主动识别和减轻安全威胁，保护敏感数据并维护 LLM 应用程序的完整性。

5.有效的成本管理：
观察 LLM 模型的资源消耗和利用情况允许组织根据实际使用模式优化资源分配和成本。可观测性工具有助于识别资源瓶颈或利用不足。这些见解可以为扩展资源或缩减资源提供决策依据，确保 LLM 应用程序的成本效益。



▲ 图 33 Datadog DCGM监控看板

4.3.2 案例2：某人工智能实验室运维实践

某国家人工智能实验室推出一系列大模型，涵盖了多个领域和应用场景，如通用大模型、气象、城市规划、工业设计等，展示了该实验室战略性、原创性、前瞻性的科学研究和技术攻关能力。其智算平台在此期间提供了强大的算力支撑能力，同时也为使用者提供了优质的运维服务，典型优势如下：

1.快速的集群交付能力：

相比自建集群，该智算平台即开即用，显著提升平台环境部署和搭建的效率；同时智算集群能够充分利用其他厂商的异构计算资源，如通用算力和超算集群算力等。

2.高效的资源调度能力：

在大规模多机多卡训练任务调度方面，智算平台提供了灵活的调度策略，如 FIFO、遍历、均衡、智能调度等，满足用户多训练场景需求，同时提升集群整体的资源利用率。

3.多样化的数据存储能力：

根据应用场景，智算平台可以选择对接对象存储、文件存储等多种存储类型；其中智算版文件存储采用分布式并行架构，基于全闪介质提供数百万级读/写 IOPS、数十 Gbps 读/写吞吐能力。

4.完善的运维运营工程化能力：

在大规模并行训练阶段，智算平台能够提供全面的日志采集、监控、即开即用的工程调优、资产、权限和任务管理等工程化能力，让使用者可以全方位通过运维指标监控、管理和优化相关计算任务。

同时从实际运维经验出发，有两类问题尤其突出，分别为资源的有效利用以及存储的合理使用：

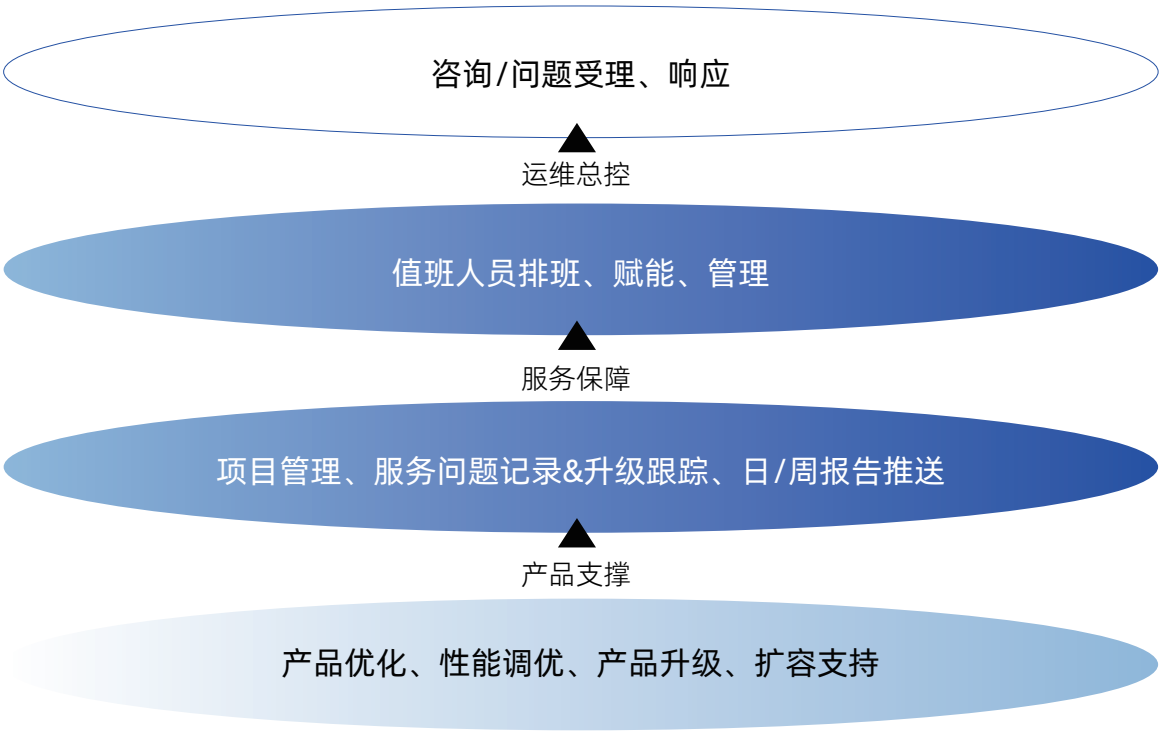
1.资源利用率优化：

多项目场景下，不同项目资源需求是动态变化的，平台应有动态调整资源的能力；同一项目下，通过资源池的子配额能力可以大大减少资源碎片的产生；多任务场景下，优秀的调度算法可以减少任务排队时间，同时也要能保障优先级较高任务的资源优先占用权等。

2.存储使用规划：

随着模型训练规模的扩大、训练数据的增加，往往需要 TB 乃至 PB 级别的存储空间来存放训练数据、模型参数、中间检查点文件等；训练过程中频繁读取数据和模型参数，以及保存训练状态，要求存储具有高速的读写能力。这需要用户充分评估模型大小、训练效率、数据吞吐 IO、数据可靠性，并充分合理搭配使用高低端存储。

为保障项目的顺利开展、平台的稳定运行，问题和需求得到及时响应和反馈，组件服务阵型如下：



▲ 图 34

05

智算平台运维运营未来展望

智算平台运维运营未来展望

智算平台的运维运营对工程能力比较薄弱的用户、运维运营团队不够完善的企业都具有重要的意义。AI 产业化对智算平台的要求不仅仅是维护大规模智算设备和调度大规模智算设备的平台，还是一个可以随开即用的、能够行业成熟算法和模型的服务。这要求技术服务可以提供数据解决方案、上云解决方案、用户和资源管理解决方案等，帮助算法工程师将AI模型调整为合适的业务模型，真正在实际业务中发挥价值。伴随技术的不断演进和用户需求的丰富，智算平台运维运营还有巨大的完善和发展空间。

1、自动化与智能化运维：

随着AI技术的进步，智算平台的运维将更加依赖自动化和智能化工具。例如使用机器学习算法预测系统故障、自动优化资源分配、智能异常检测与自我修复能力，AIOps 也将深度学习、大数据分析等技术融入传统运维流程，实现对大规模集群的高效管理。自动化与智能化技术的应用将大幅减少人工干预，替换部分重复性的运维运营工作，提高运维效率。

2、持续集成/持续部署（CI/CD）：

在 AI 开发和部署流程中，CI/CD 加速模型从研发到生产的流转速度，同时保证代码质量与版本控制，未来智算平台也应与 CI/CD 技术及工具深度融合，打造一站式的 AI 开发部署平台。

3、运维运营人员的培养：

建立更成熟的运营运维人员培养体系，产出标准化的培训课程，搭建相关的实训体系，优化运维运营专业人员紧缺的现状，优化智算生态专业人员构成。

智算平台的运维运营正朝着自动化、智能化的方向发展，为AI的规模化应用打下更加坚实的基础。