构建云原生算力基础设施 驱动大模型创新实践

王羽中　杭州谐云科技有限公司

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

演讲嘉宾

王羽中

杭州谐云科技有限公司 技术总监

负责谐云科技边缘智能、算力管理平台、MLP平台、大模型支撑平台等核心产品的技术演进。研究关注范围包括云原生技术、AI大模型、云边协同技术等，具有丰富的大规模底层支持系统架构设计经验和实践落地经验。

目录
CONTENTS

PART 01
背景介绍

大模型的重要性

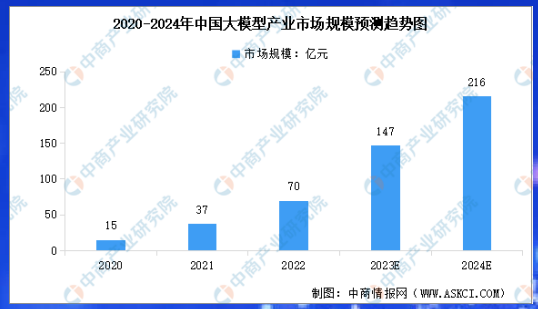
自2022年11月30日ChatGPT发布以来，AI大模型在全球范围内掀起了有史以来规模最大的人工智能浪潮。大模型因其拥有表达能力好、泛化能力好、能够处理复杂任务和语义理解、知识库存储容量大等优势很快迎来了迅猛发展。大模型将重新塑造人类知识应用、创造和转化的模式，在经济社会发展中产生巨大价值。

政策方面

- 2022年8月，科技部发布《关于支持建设新一代人工智能示范应用场景的通知》中指出坚持面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康，充分发挥人工智能赋能经济社会发展的作用，打造形成一批可复制、可推广的标杆型示范应用场景。
- 2023年12月，国家数据局发布《“数据要素x”三年行动计划（2024—2026年）（征求意见稿）》，提出以科学数据支持大模型开发，建设高质量语料库和基础科学数据集，支持开展通用人工智能大模型和垂直领域人工智能大模型训练。同时，北京、上海、深圳、安徽、四川等省市也陆续出台大模型产业发展措施，加速大模型应用落地。
- 其他政策.....

产研方面

- 中商产业研究院分析师预测，2023年中国大模型产业规模将达到147亿元，2024年将达到216亿元。
- 据不完全统计，截至2023年8月，中国已发布的各类大模型数量已超上百个；根据科技部“新一代人工智能发展研究中心”发布的数据，截止至2023年我国参数规模在10亿以上的大模型总数量达79个。截至2023年10月，拥有10亿参数规模以上大模型的厂商及高校院所达到了254家。



数据来源：中商产业研究院整理

大模型落地情况

统计数据显示，在大模型落地应用中，45%的企业处于观望阶段、39%的企业处于探索可研阶段、16%的企业处于试点应用阶段，而全面落地应用的企业为零。

大模型落地面临的关键问题：

- 大模型幻觉

现阶段，大模型输出准确度能够达到70%-90%左右。由于对准确性、可控性要求较高，大模型面客应用都暂时无法落地。应用将以对内为主。

- 答案时效性

大模型需要将最新数据通过预训练方式灌注到模型中，因此很难囊括最新知识，其回答内容的时效性也受到限制。

- 数据质量问题

当前专业的领域知识数据都孤立在各个企业和机构中，没法构建大规模高质量的数据集，造成大模型在专业领域和垂直行业效果不佳。

国内外大模型对比

2019-2023年1-5月全球AI大模型数量对比

	2019年	2020年	2021年	2022年	2023年1-5月
美国	4	11	30	37	18
中国	0	2	30	28	19
其他	0	1	8	11	1

SuperCLUE总排行榜（2024年2月）

排名	模型名称	机构	总分	OPEN	OPT	使用	发布日期
-	GPT4-Turbo-0125	OpenAI	92.71	94.95	83.74	API	2024年2月27日
-	GPT4（网页）	OpenAI	90.36	91.48	85.89	网页	2024年2月27日
🥇	文心一言4.0	百度	87.75	88.23	85.82	API	2024年2月27日
🥈	GLM-4	清华&智谱AI	86.77	87.49	83.89	API	2024年2月27日
🥉	通义千问2.1	阿里巴巴	85.7	86.1	84.09	API	2024年2月27日
4	Baichuan3	百川智能	82.59	82.45	83.13	API	2024年2月27日
5	Moonshot(KimiChat)	月之暗面	82.37	82.29	82.66	网页	2024年2月27日
6	讯飞星火V3.5	科大讯飞	81.01	80.6	82.64	API	2024年2月27日
7	qwen1.5-72b-chat	阿里巴巴	79.36	78.39	83.22	API	2024年2月27日
8	MiniMax_Abab6	稀宇科技	78.01	77.94	78.29	API	2024年2月27日
10	从容大模型V1.5	云从科技	75.56	75.69	75.03	API	2024年2月27日
9	云雀大模型	字节跳动	76.58	75.47	81.04	API	2024年2月27日

- 在大模型数量上，我国已经和美国逐年持平；
- 在模型的效果上依旧存在较大的差距；据专业的SuperCLUE组织公开数据显示，截止2024年2月，美国以GPT-4为代表的大模型的总分达到92.71，国内以文心一言4.0为代表的大模型的总分为87.75；
- 国外以GPT-4为代表的大模型参数规模已经达到了1.8万亿，国内以文心一言4.0为代表的大模型参数规模尚未突破万亿规模的参数，参数规模是影响模型效果的重要因素之一；

算力对大模型的重要性

Large Language Models: A New Moore's Law?》——Julien Simon@Hugging Face

- 在大模型领域，模型的性能通常与其规模成正比。也就是说，模型越大，它的性能和表现就会越好。
- 作为大模型的基础“底座”，算力在其中发挥着关键的作用。动辄百亿甚至千亿数据规模的大模型训练，例如OpenAI训练GPT-4，在大约25000个A100上训练了90到100天。百度文心一言4.0大模型也是在万卡的集群中训练数十天才完成。

算力规模决定大模型参数规模，从而间接决定大模型的效果和落地实践的可行性。如何解决大模型落地进程慢，缩短国内大模型与国外大模型性能差距，算力作为大模型的基础设施都起着决定性作用。要实现大模型的弯道超车，要实现大模型的全面落地实践，算力基础设施建设是重中之重。

人工智能大模型的快速发展，让算力引发前所未有的关注。伴随算力发展规划政策相继出台，算力整体布局持续优化，全国上下已形成积极推动算力产业快速健康发展的局面。

算力基础设施建设

近年来，围绕加快算力基础设施建设应用，我国出台一系列重要政策举措，实施一大批重大工程项目。截至目前，从计算设备侧看，我国近六年累计出货超过2091万台通用服务器，82万台AI服务器，算力总规模达到302EFlops，全球占比33%，增速达50%，其中智能算力保持稳定高速增长，增速达72%。

2022年全球计算指数评估报告得出，计算力指数平均每提高1点，国家的数字经济和GDP将分别增长3.5‰和1.8‰

伴随算力经济的发展，算力技术和人工智能的融合创新让智能计算中心成为新基建热点，即专门用于人工智能计算的中心。截至2023年3月，国内有超过30个城市正在建设或提出建设智能计算中心。


中华人民共和国中央人民政府
www.gov.cn

首页 > 政策 > 国务院政策文件库 > 国务院部门文件

标　题：关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见
发文机关：国家发展改革委 国家数据局 中央网信办 工业和信息化部 国家能源局
发文字号：发改数据〔2023〕1779号
来　源：国家发展改革委网站
主题分类：工业、交通\信息产业（含电信）
公文种类：意见
成文日期：2023年12月25日

国家发展改革委等部门
关于深入实施“东数西算”工程
加快构建全国一体化算力网的实施意见
发改数据〔2023〕1779号

各省、自治区、直辖市及计划单列市、新疆生产建设兵团发展改革委、数据局、网信办、工业和信息化主管部门、通信管理局、能源局：

算力是数字经济时代的新型生产力。算力网是支撑数字经济高质量发展的关键基础设施，可通过网络连接多源异构、海量泛在算力，实现资源高效调度、设施绿色低碳、算力灵活供给、服务智能随需。为深入贯彻落实党中央、国务院决策部署，加快构建全国一体化算力网，以算力高质量发展支撑经济高质量发展，制定本实施意见。

近年来，我国以“东数西算”工程为牵引，加快推进信息基础设施建设，提高算力对人工智能、数字经济等的支撑能力，助力经济高质量发展。

北京市经济和信息化局北京市通信管理局关于印发《北京市算力基础设施建设实施方案（2024—2027年）》的通知中重点指出大力推动人工智能大模型与自主可控芯片开展适配，提高我国智算产业供应链安全性、稳定性和坚韧性。

算力基础设施建设进程加快，如何建设高效、灵活、稳定的算力管理平台，向下实现算力资源的统一纳管，向上为大模型提供算力服务，加速大模型落地实践。

PART 02
关键要素

支持异构算力调度

过度依赖算力芯片进口，依然成为“卡脖子”技术

《中国算力白皮书（2022）》和中国信通院的数据，2021年第四季度，英特尔占据了全球84%的CPU算力芯片市场份额和71%的FPGA算力芯片市场份额，**英伟达占据了全球95.7%的GPU算力芯片市场份额**。

美国政府制裁分为三个层面：

1、美国要求英伟达和AMD停止对华供货高端GPU。

2、限制芯片设计人才在华就业。

3、限制为大陆芯片企业代工。

完善算力综合供给体系

- 优化算力设施建设布局
- 推动算力结构多元配置，推动不同计算架构的智能算力与通用算力协同发展。

工业和信息化部等六部门联合印发《算力基础设施高质量发展行动计划》

算力国产化取得显著成果

据统计，近年来国产芯片自给率不断提升，2019年为30%《中国制造2025》计划要求在2025年，国产芯片自给率要达到70%以上。

Ascend　Cambricon 寒武纪科技　HYGON 中科海光

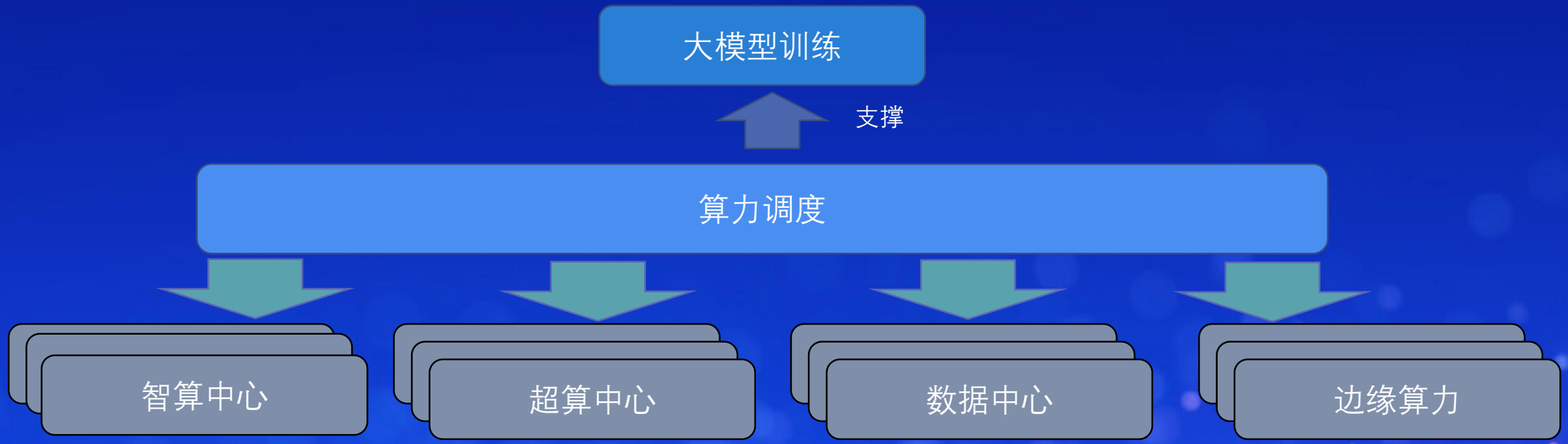
登临科技　昆仑芯 KUNLUNXIN　Enflame 燧原科技

摩尔线程 MOORE THREADS　壁仞科技 BIREN TECHNOLOGY　天数智芯 Iluvatar CoreX

未来的算力中心必定是存量的英伟达GPU和国产的AI芯片共存的模式，因此算力管理平台必须能统一纳管英伟达GPU和国产AI芯片等各种异构算力资源，实现算力的统一分配和调度；

支持跨算力中心调度

- 动辄百亿、千亿甚至万亿参数规模的大模型训练需要大量的算力资源支撑，例如OpenAI训练GPT-4，在大约25000个A100上训练了90到100天。百度文心一言4.0大模型也是在万卡的集群中训练数十天才完成。
- 国内缺乏超大规模算力集群，单算力中心算力资源有限，无法支撑超大规模大模型的高效训练；
- 智算中心、超算中心、数据中心、边缘侧都分布有算力资源，无法有效整合，造成资源浪费；



算力管理平台需要支撑跨算力集群和算力中心的算力管理和调度；有效整合分散在各个算力中心的算力资源，聚少成多，为大模型的训练提高算力支撑；

支持多种类型任务调度

大模型和小模型的结合将成为未来AI产品的重要发展趋势，也是人工智能应用赋能行业发展的重要方向。

大模型优势

- 拥有更多的参数，能够更准确地捕捉数据中的模式和特征，处理复杂任务的表现更好，能够实现更准确、自然的内容输出；
- 通过学习大量数据中的细微差异，能够更好地适应任务需求，在处理大规模数据集或未见样本的预测表现更出色；
- 大模型能够处理更复杂的语言结构，理解更深层次的语义；
- 拥有更大的容量，可以存储更多的知识和经验；

小模型优势

- 参数量较少，因此训练和推理速度更快；
- 占用资源较少，小模型在移动设备、嵌入式系统或低功耗环境中更易于部署和集成，占用资源少，能够在资源受限的设备上运行；
- 当面对少量标注数据时，大模型可能会因为过拟合而出现性能下降的情况，而小模型通常能够更好地泛化，提供更准确的结果；
- 在一些特定场景下，效果反而比大模型更好；

算力管理平台需要支持大模型、小模型等多种类型任务的调度；

支撑算力精细化调度

2018年，AWS在AWS re:Invent2018大会曾提及，在AWS上GPU利用率只有10%至30%。

2020年，香港IDC新天域互联公布数据，企业通常仅使用15%至30%的GPU服务器资源。

原创 GPU服务器算力神通 但利用率仅15%什么情况?

2020-06-22 09:39

越来越多的企业正积极投资于机器学习（ML）和深度学习（DL）应用，这些程序将大量数据转换为可产生业务价值的人工智能（AI）用例。与此同时，企业需要加强基础设施的运作效率，以加快训练和驱动AI的机器学习和深度学习模型的过程。

在寻找专用于超高计算性能基础方案时，企业最常选用GPU服务器租用，值得注意的是，企业应该时刻思考所付出的投资是否能够全部价值化。根据香港IDC新天域互联获悉的一份GPU服务器利用率报告，显示企业通常仅使用15%至30%的GPU服务器资源，虽然某些GPU服务器存在共享情况，但绝大多数仍是针对单个用户。

算力管理平台对任务的精细化调度：

- 算力聚合和单卡共享；
- 算力超分和优先级调度；
- 算力动态分配和调度；

实现算力资源利用率的有效提升，发挥算力资源的最大价值，实现降本增效。

其他关键要素

开箱即用的算力服务；

提供一些算力模版和服务套件，支持用户快速使用算力服务；

支持对内提供算力服务，对外提供算力运营；

算力中心的算力在部分时间处于空闲状态，支持算力对内提供算力服务，对外提供算力运营；

算力的精细化计量计费；

在算力运营场景下，支持算力的精细化计量计费，支持包年包月和按量计费等模式，支持任务级、秒级的计费粒度；

算力的统一监控运维；

支持对算力中心和算力集群的统一监控运维，管理员可以很方便的了解到各个算力中心的健康状态、资源情况、任务运行情况等信息；

算力的租户隔离；

在多组织、多租户的使用场景下，支持算力在租户间的分配、限制和隔离；

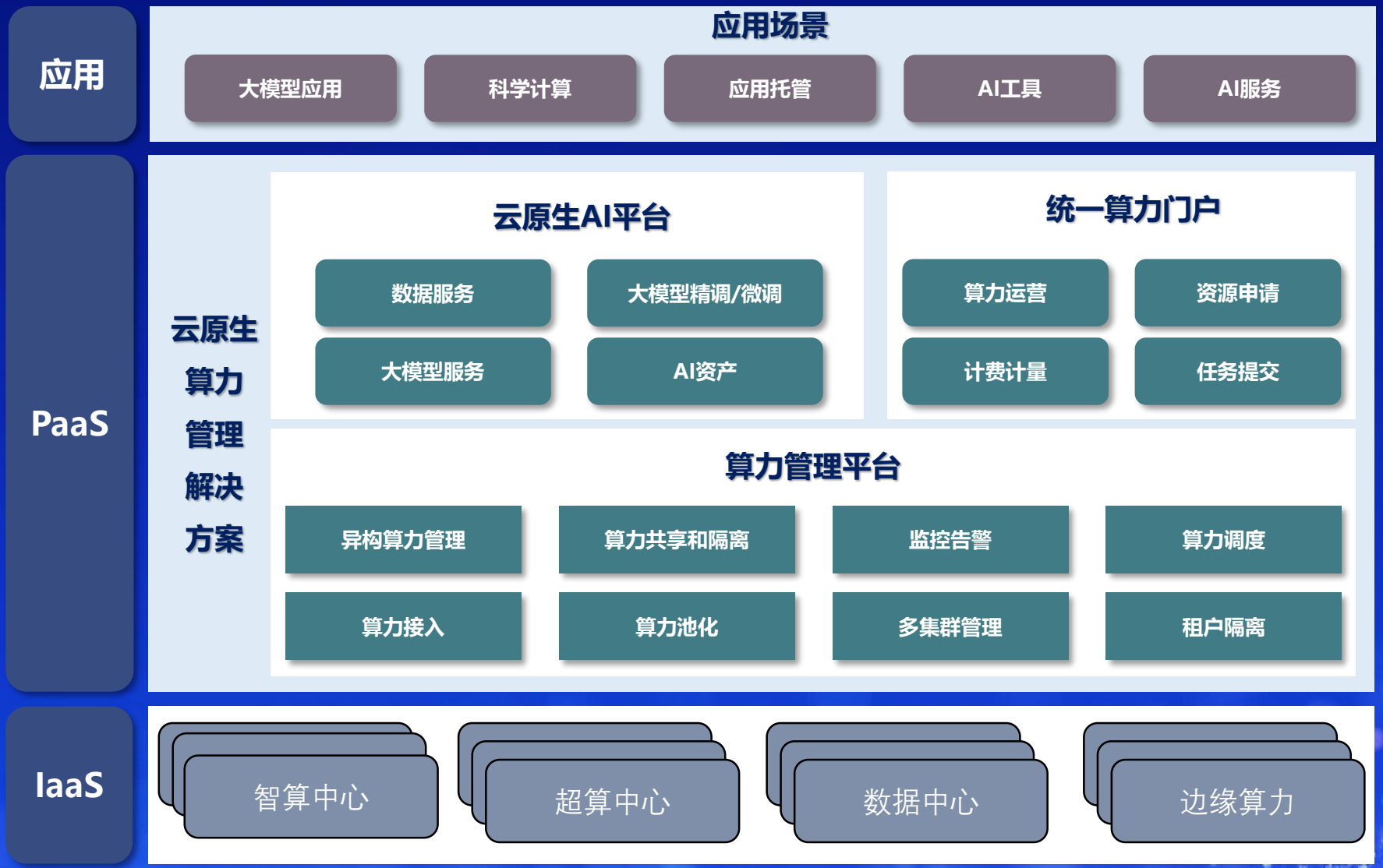
PART 03
技术方案

云原生是建设算力管理平台的最佳方案

资源纳管	以Kubernetes为代表的云原生技术支持大规模资源、异构资源的高效管理和运维，也提供了灵活的扩展方式。
应用支撑	云原生技术的焦点就是支撑分布式、微服务等应用的编排调度、弹性扩所容、高可用等，因此针对大模型场景下精调/微调任务，大模型服务等都具有很好的支撑。
生态成熟	云原生技术经过多年的发展已经趋于成熟，在日志监控、权限控制、租户管理等方面都有成熟的生态系统支撑。
主流趋势	云原生技术作为一种公认的技术趋势已经广泛被用于算力基础建设，据统计90%以上的智算中心都采用云原生技术建设。

云原生是算力基础设施建设的核心技术，是发挥算力资源效能的最佳实践路径。

平台架构



统一算力门户

基于算力管理底座提供算力资源申请、运行、监控等管理全流程。

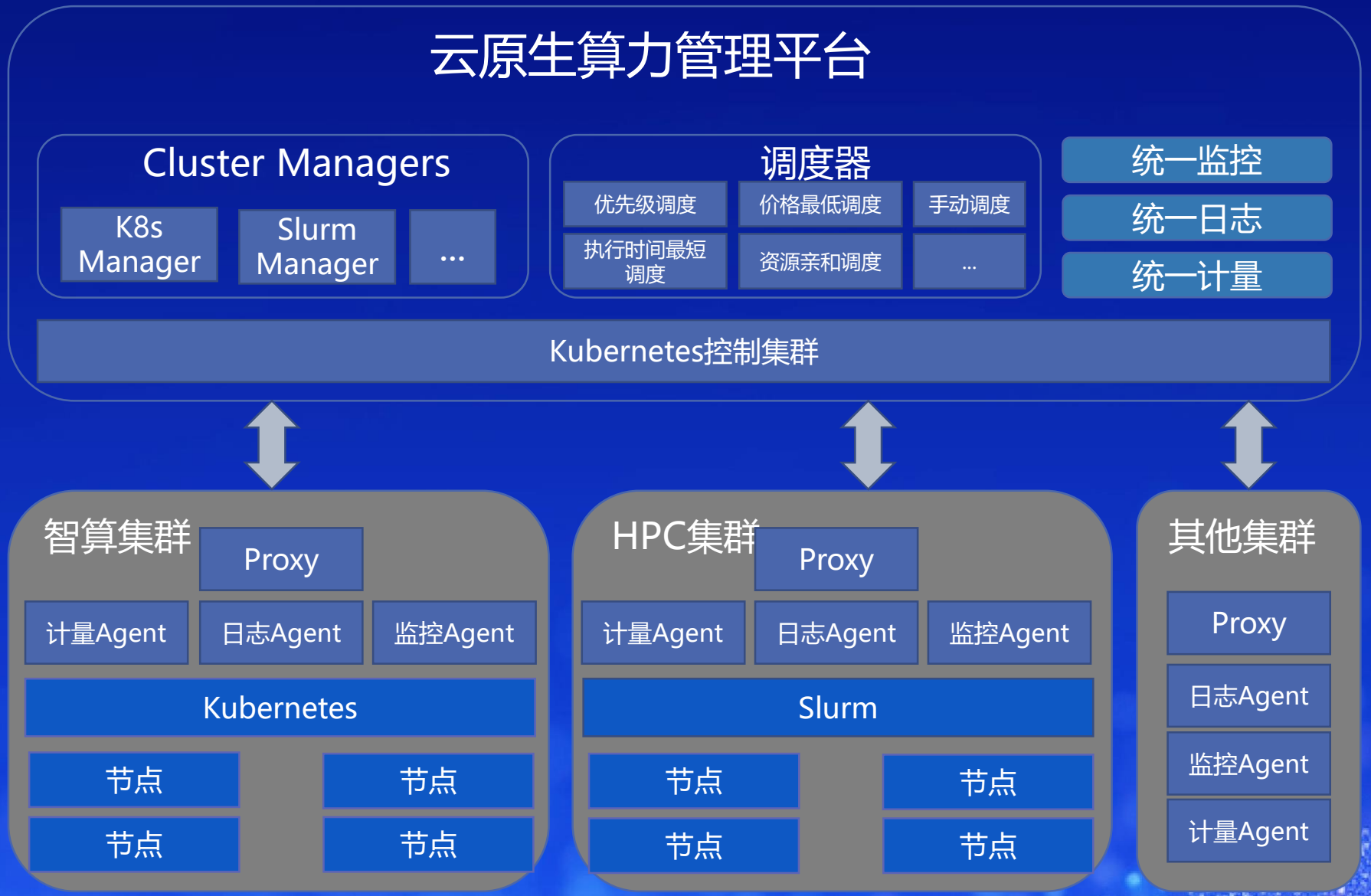
云原生AI平台

针对大模型精调/微调、大模型服务等场景，构建云原生AI平台，整合分布式训练、服务部署、数据服务等能力，提高大模型开发部署效率。

算力管理平台

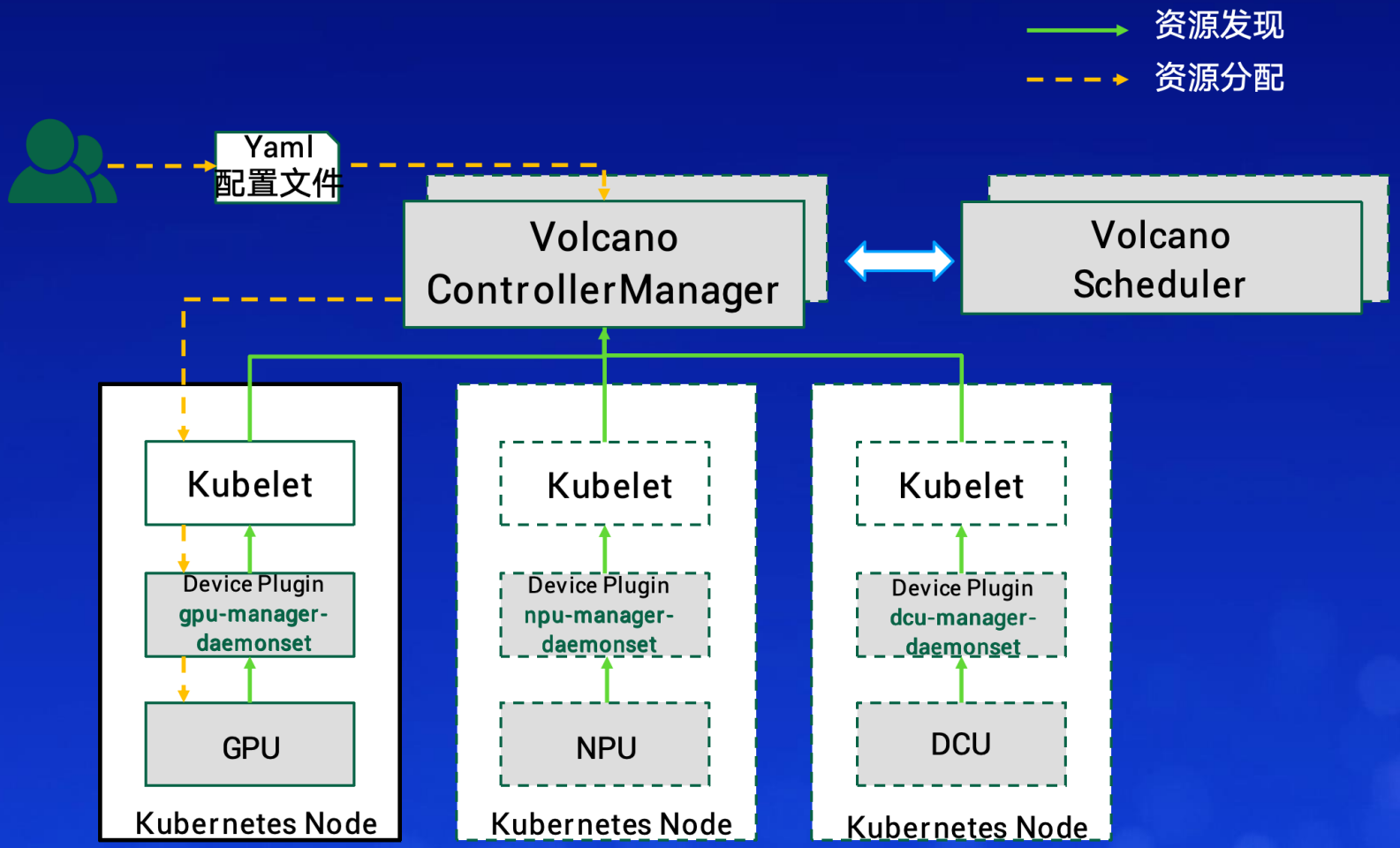
针对智算算力、HPC算力、边缘算力等算力资源，非侵入式接入异构资源，通过按需分配、精细化管理与调度，为大模型应用、算力运营等提供算力底座支撑。

关键技术点1-跨算力中心的纳管和调度



- 上层控制集群中引入自研调度器，实现将用户提交的任务调度到对应的算力集群中运行，支持优先级调度、价格最低调度等多种调度策略；
- Proxy和Manager一一对应，实现任务的下发和底层算力集群的状态、资源、任务状态等上报；
- 针对不同类型集群开发对应的Proxy，实现底层集群的差异性屏蔽；智算集群的Proxy与API Server交互，HPC集群的Proxy与Slurm交互，一些自带管理系统的算力集群（商业系统、公有云等）的Proxy与管理系统的API交互；

关键技术点2-异构资源纳管和调度

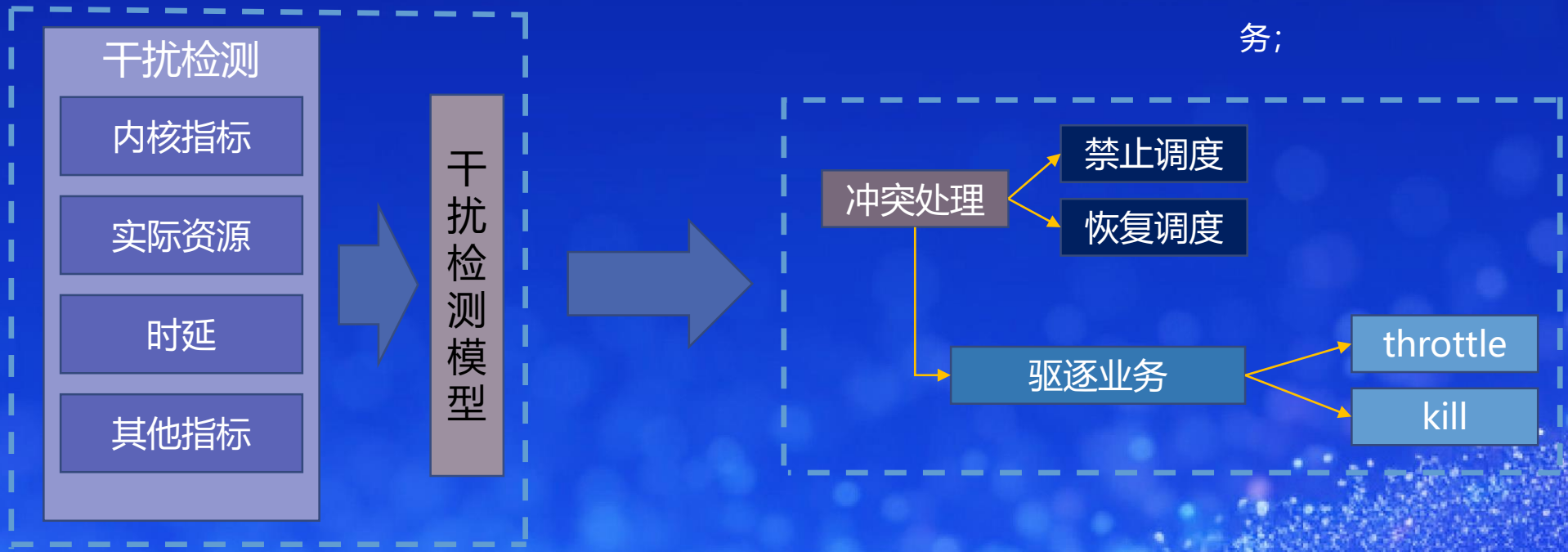


- Device Plugin注册、管理AI算力卡在社区已经成为事实标准。基于Device Plugin实现各种异构资源无侵入的注册和接入；
- 基于Volcano的高性能工作负载调度引擎实现AI、高性能计算等批量计算任务调度和编排管理和调度。
- 调度引擎支持按卡类型调度、资源空闲最多调度、binpack调度、批量调度、资源均衡调度等多种调度策略；
- 基于Queue实现租户间的资源隔离和资源限制；

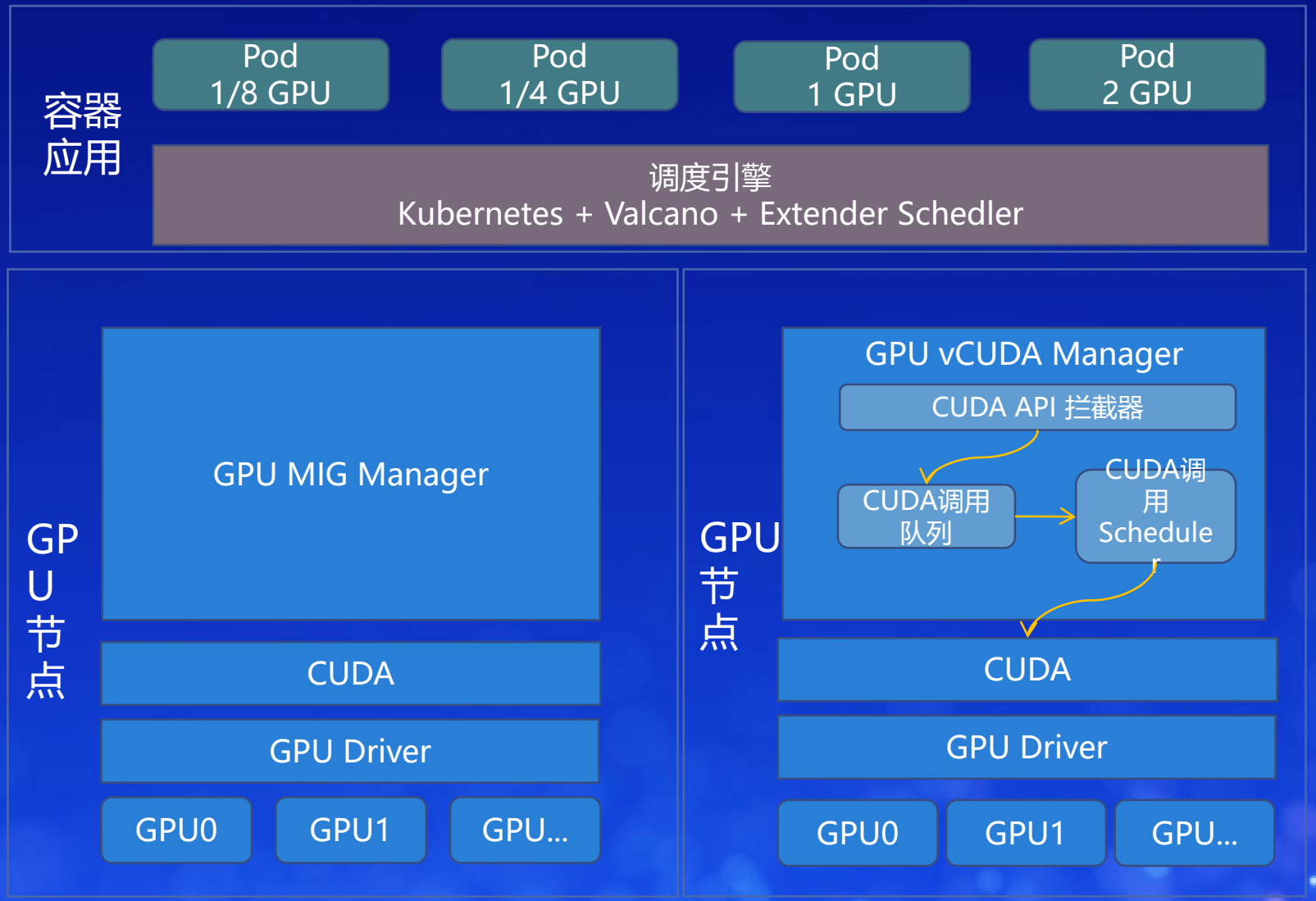
以云原生技术为核心实现对多种异构资源的统一管理与调度。

关键技术点3-算力超分和优先级调度

- 支持资源超分，所有队列申请资源总和可大于集群实际资源总和；
- 基于任务的资源实际使用情况和资源预测，动态计算和调整高低队列资源大小；
- 当高优先级队列提交的任务没有足够资源运行时，可以驱逐和抢占低优先级队列资源；
- 构建干扰检测模型实时监测高优先级任务是否受到干扰，高优先级任务受到干扰时，可以压制和驱逐低优先级任务；



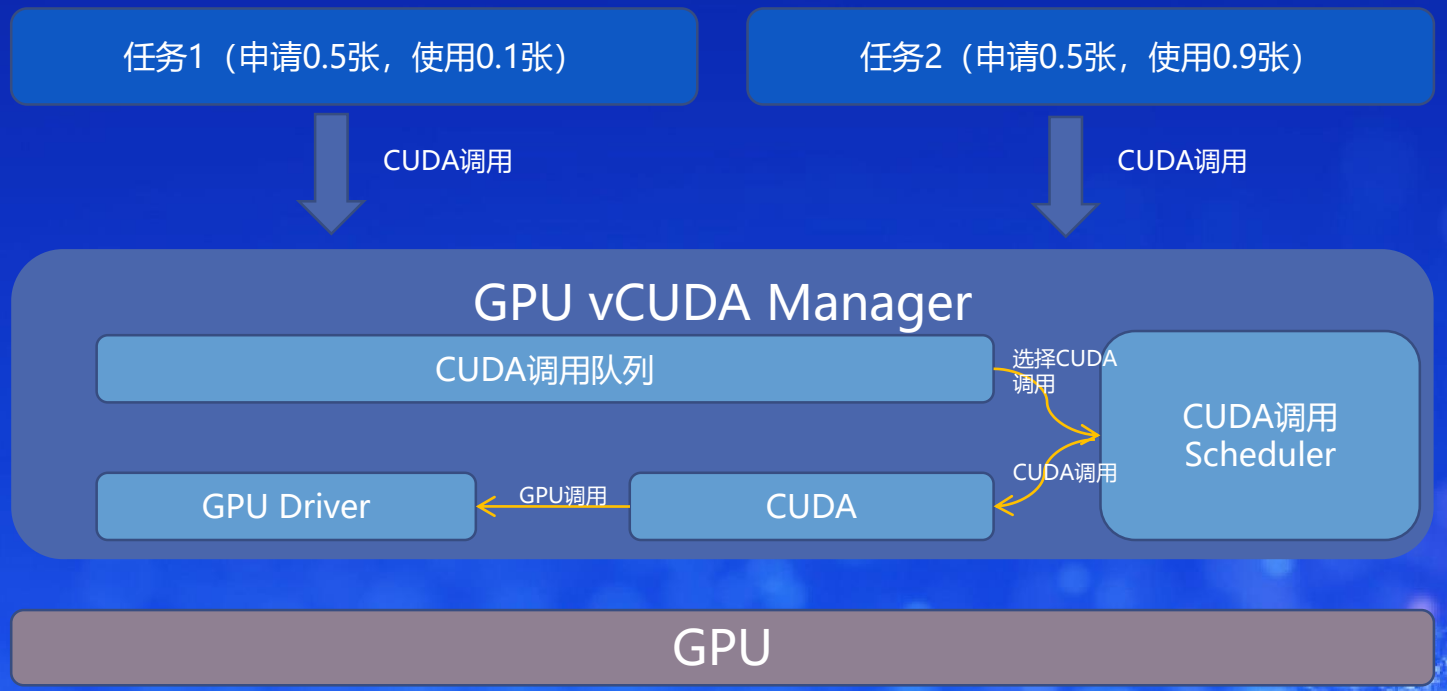
关键技术点4-算力资源共享和隔离



- 支持MIG和vCUDA两种算力资源共享和隔离方式；用户可针对业务场景和应用类型灵活选择不同的方式；
- 在CUDA调用层针对业务编程接口进行适配改造，实现算力和显存的调度API拦截，并基于Schuedler实现调用频率的控制，从而实现算力细粒度共享和隔离。最小支持以 0.01算力，1MB显存的 vGPU 供业务使用，透明无感。这种方式不依赖于各个硬件厂商的能力，便于扩展。
- MIG支持SIngle和Mixed两个模式的动态配置和调度。

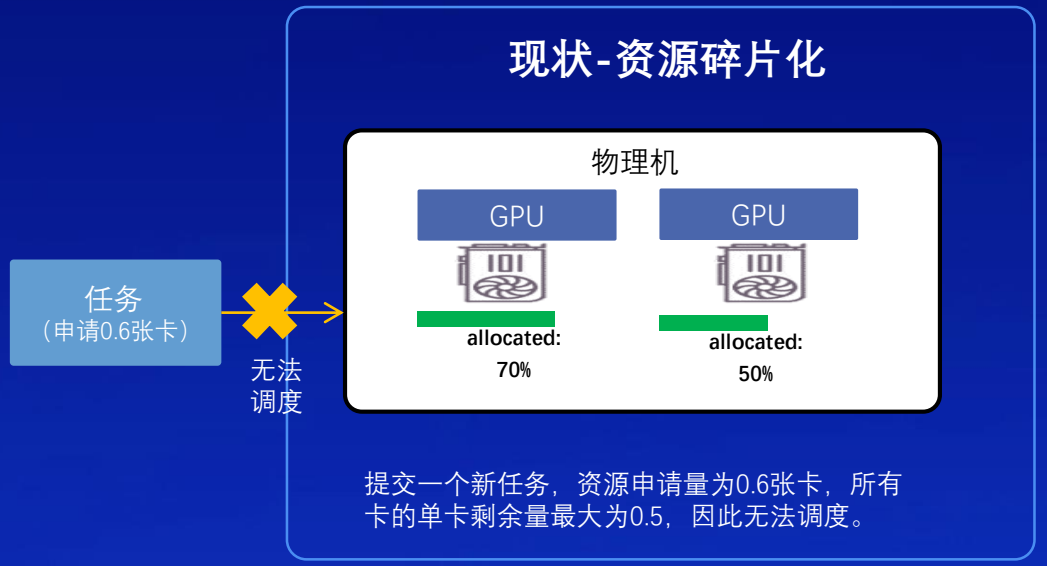
关键技术点5-算力资源动态共享

- 资源静态分配，有些任务申请的资源大部分时间处于空闲状态；
- 某些任务在峰值时需要的算力资源会超过申请的量，但是最多只能使用申请的算力资源；
- 算力资源利用率底，但部分任务又存在资源不足的情况；

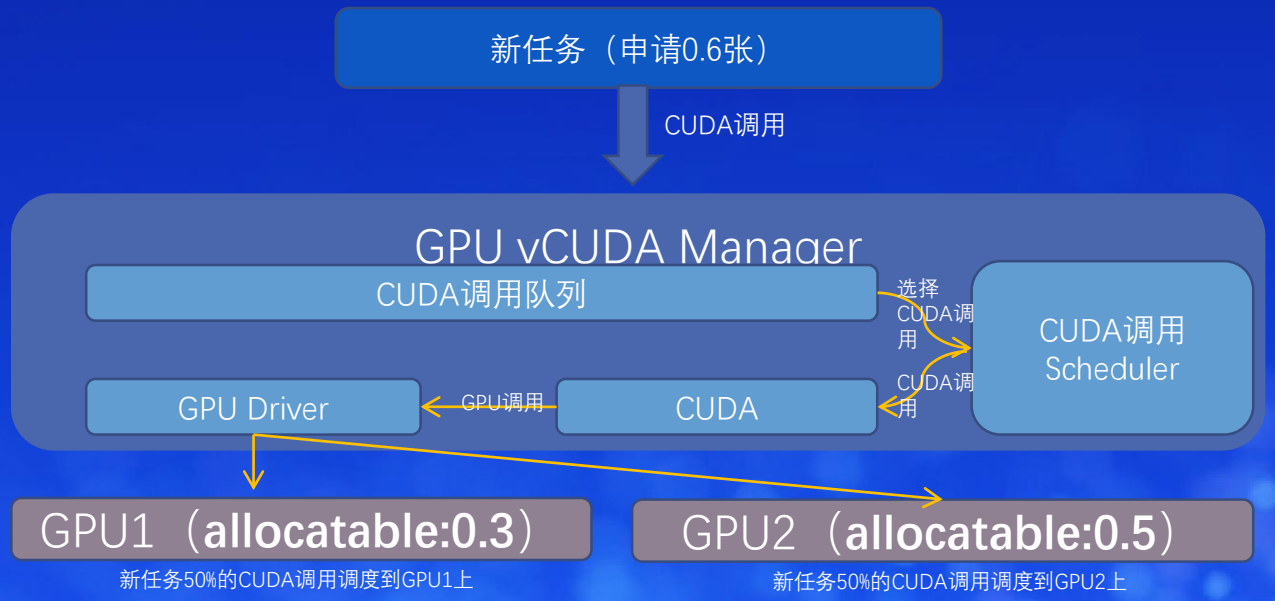


- 容器可使用的算力资源可根据共享同一张卡的其他任务容器的使用情况动态调整；
- 引入limit特性，限制使用上限；
- 引入节点算力资源超分能力，即节点上的容器的算力资源申请总和可大于节点实际资源，CUDA调用Scheduler可根据任务的优先级选择CUDA调用，在资源发生抢占时，优先调度高优先级任务的CUDA调用，保障高优先级任务的资源需求；

关键技术点6-多卡共享



- 在算力卡共享的场景下，经常存在算力卡资源碎片化问题；
- 如果单卡的剩余资源量无法满足新任务的算力资源申请量，新任务将无法调度运行；
- 存在整台物理机资源很充足，但新任务无法调度的现象，造成算力资源利用率低；



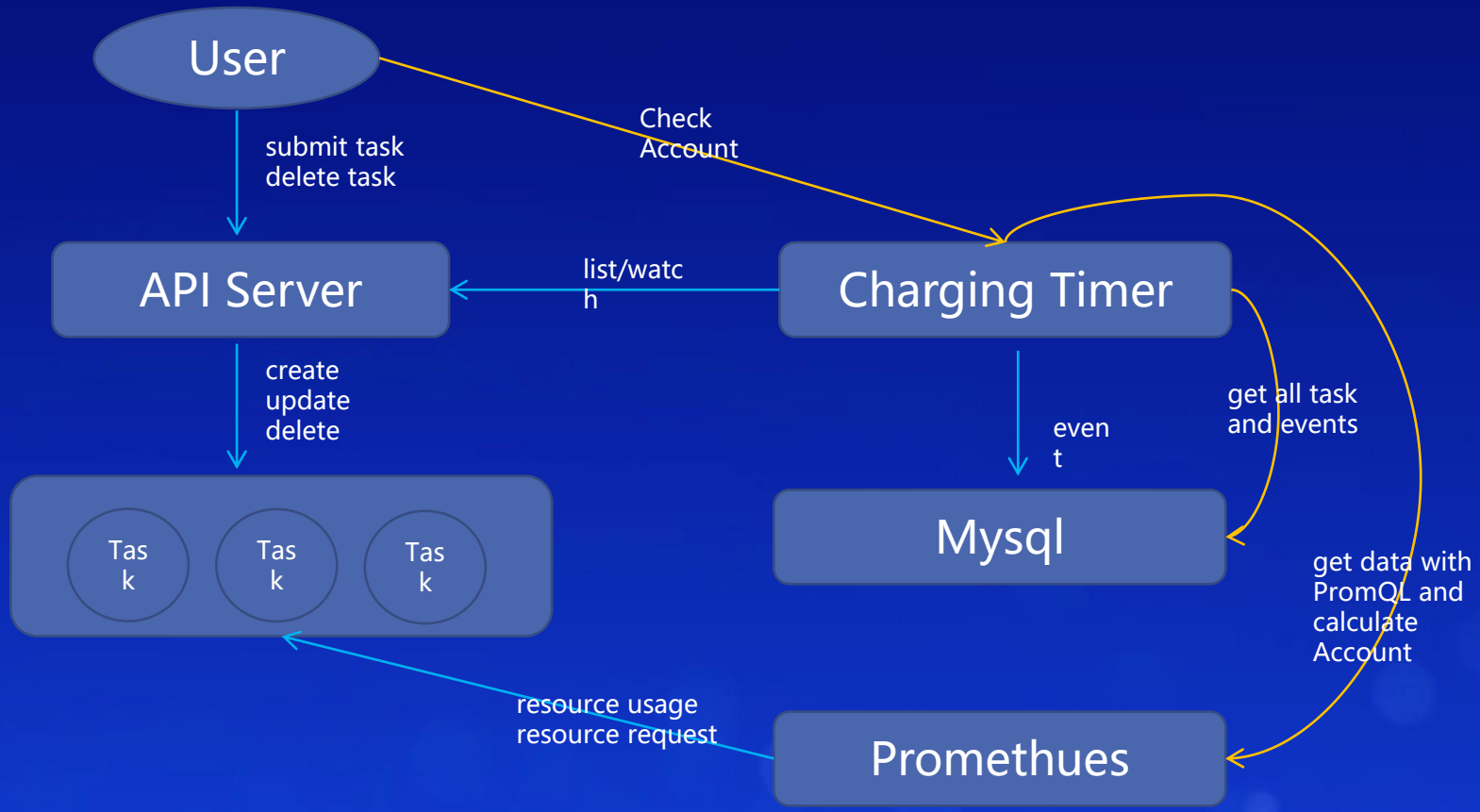
- GPU vCUDA Manager根据节点上每张卡的可分配量以及新任务的算力资源申请量，自动寻找最优的调度方案；
- 如果有单卡能满足新任务的算力资源申请量，优先将新任务调度到单卡上，如果没有单卡能满足新任务的算力资源申请量，则以碎片化最小原则为调度目标将任务调度到多张卡上；
- CUDA调用Scheduler根据调度方案，将对应比例的CUDA调用调度到对应的算力卡上；

关键技术点7-精细化计费计量

支持计费模式：
包年包月（裸金属、虚拟机）；
按使用量计费（容器）；

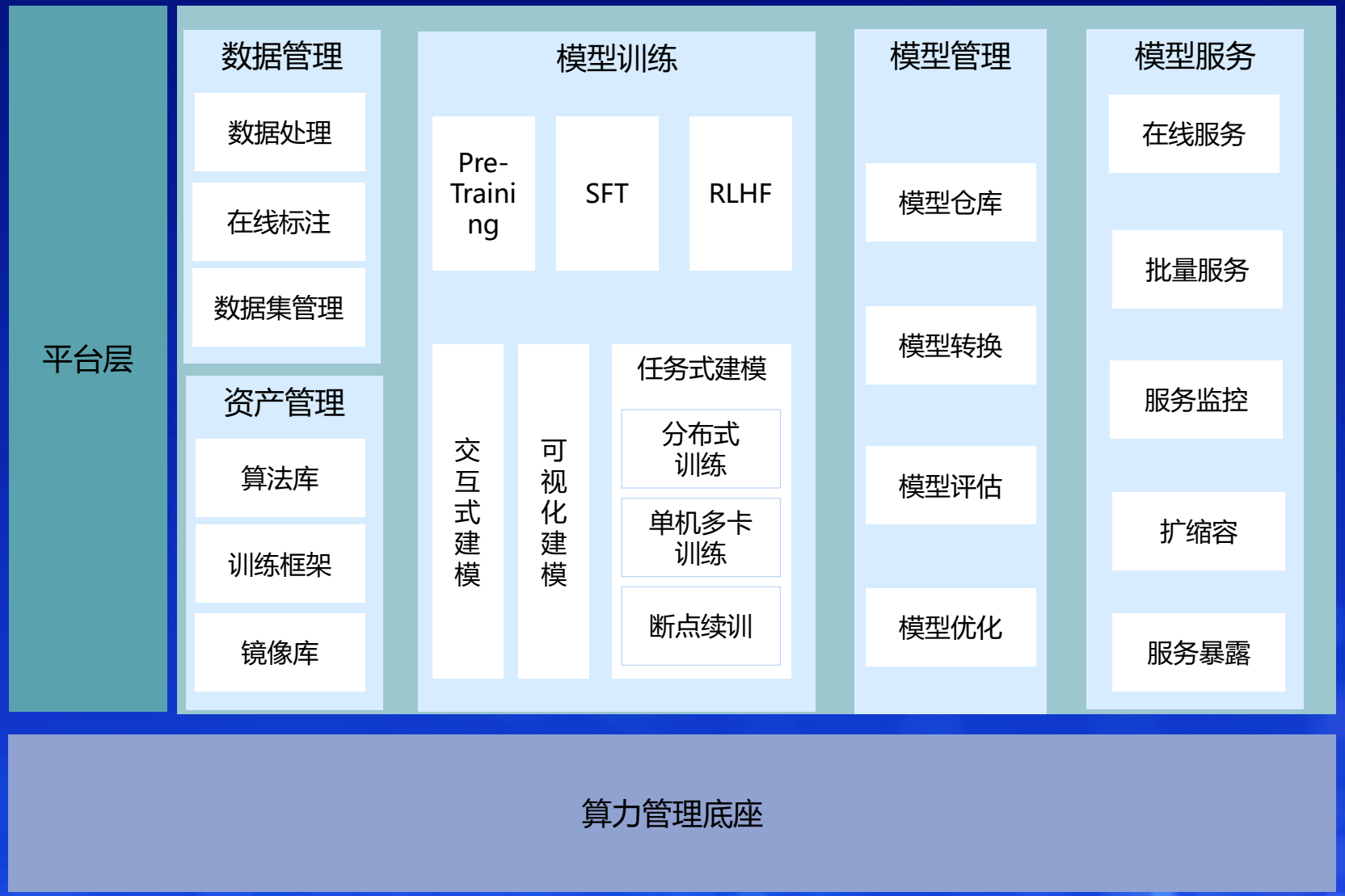
支持的workload类型：
- argo workflow；
- volcano job；
- k8s deployment；
- k8s job；

支持统计的资源纬度：
- CPU；
- Memory；
- GPU；
- GPU Memory；
- NPU；
- NPU Memory；
- Storage；



- Charging Timer通过list/watch实现对所有任务的提交、运行、删除等用户行为感知，并记录相应的Event（User、Task、Action、Time）到数据库中；
- 当用户查询账单时，先查询指定时间段内的所有任务和任务事件，并拼接成PromQL语句去Promethues中查询任务的资源使用数据，最终计算出资源使用量数据和账单数据；
- 针对容器的爆发模式，使用Request和Usage取最大值的方式实现精确计费；
- Charging Timer模块会按天进行账单合并，增加账单查询速度；

关键技术点8-云原生AI平台



一站式大模型应用开发上线体验

支持从数据上传、数据预处理、数据标注、模型训练、模型评估到模型部署发布的大模型开发上线全流程；

丰富的模型训练方式

支持交互式建模、可视化建模、任务式建模等多种训练方式，满足各种场景、各种用户的需求；

支持主流分布式训练框架

支持TensorFlow PS、PyTorch DDP、MPI等主流的分布式训练框架，支持断点续训等能力；

一键式的大模型精调/微调

平台集成常用的开源大模型，支持用户基于领域数据一键式完成大模型精调/微调；

典型案例

某高校面向校内和校外的科研需求建设一个国内领先的算力中心，谐云为该高校计算中心打造的异构资源管理平台，统一管理高校自有算力中心与来自各类运营商、云厂商等提供的算力资源，实现资源一站式管理与运营，提升用户体验。

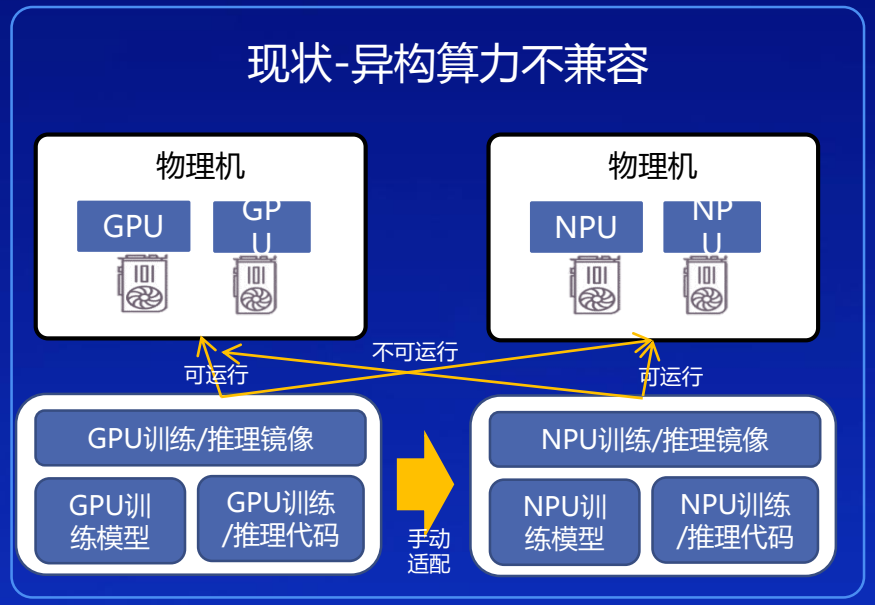
典型案例

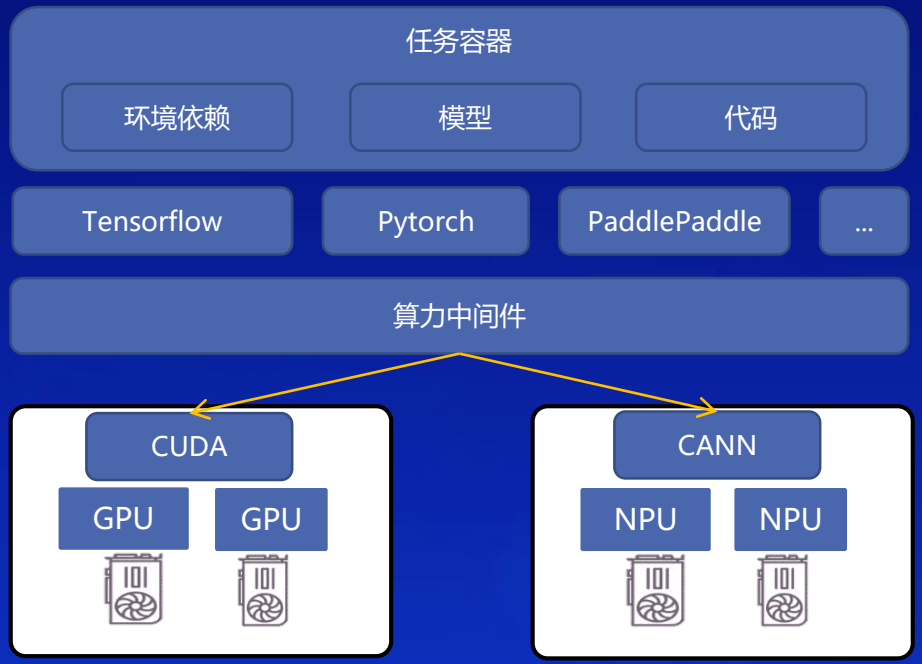
某政府针对算力资源、数据、行业算法模型等资源分散、无法高效利用等问题，联合谐云建设一体化MLP平台，实现：

- 实现对全市异构的、分散的算力进行统一纳管，并对外为用户提供算力服务；
- 提供大模型、小模型等全链路服务，从数据标注-模型构建-模型训练-模型服务-服务应用的端到端服务；
- 提供将数据、模型、算法等进行共享，提高资源复用能力；

PART 04
未来展望

算力中间件





- 跨算力类型的镜像、模型和代码都要做相应的适配，工作量极大；
- 无法实现真正意义上的异构算力调度，用户在运行任务时必须指定运行在哪种类型算力资源上；
- 异构算力无法真正融合使用，单一类型算力资源有限，无法支撑大规模任务的运行；

- 研发算力中间件，实现算子指令的自动转换，屏蔽底层的异构算力，真正做到同一个任务容器可以跨算力类型运行；实现真正意义上的异构算力调度；
- 实现异构算力的池化，支撑大规模任务的运行；

跨集群算力池化和调度

现状	• 当前单个应用只支持在单个算力集群中运行； • 算力集群之间的资源孤岛问题依然存在；

- 基于分布式框架将分布式任务切分成多个Task，并调度到不同的算力集群中运行；
- 通过平台调度引擎实现单一分布式任务跨算力集群调度，实现分散算力集群资源池化，支撑超大规模任务执行；

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

THANKS