HUAWEI

智驾算力运营白皮书



目录

CONTENTS

PART 1

智驾算力运营白皮书编制背景与目的 01

1.1 智驾算力应用现状 01

1.1.1 智驾算力迈入超大规模集群化时代 01

1.1.2 智驾算力向多元异构方向演进 02

1.1.3 智驾研发呈多模型、多范式、多任务并行态势 02

1.2 智驾算力运营面临的核心困难与挑战 02

1.2.1 超大规模算力扩张下的供需错配挑战 02

1.2.2 多元异构算力统一度量与运营挑战 02

1.2.3 多任务并行研发下的资源分配与协同挑战 03

1.2.4 复杂场景下的效能管控与成本平衡挑战 03

1.3 智驾算力运营白皮书编制目的 03

PART 2
算力运营机制 04

2.1 资源规划管理机制 04
2.2 资源申请开通机制 05
2.3 任务与优先级管理机制 05
2.4 资源效率管理机制 06
2.5 周报/月报、季度复盘机制 06

PART 3
算力运营指标体系 07

3.1 智驾算力运营指标体系的构建原则与核心场景 07
3.2 智驾算力运营指标体系的分层结构 08
3.3 智驾算力运营基于四层架构的核心指标参考 08

PART 4
算力运营平台建设 14

4.1 平台参考架构 14
4.2 核心功能模块说明 16

PART 5
典型算力运营优化策略及案例 17

5.1 典型算力优化目标及策略 17
5.2 算力运营优化案例 18

PART 6
算力运营成熟度模型 22

6.1 构建算力运营成熟度模型的意义 22
6.2 算力运营成熟度等级定义 23
6.3 算力运营成熟度应用建议 24

PART 7
行业倡议与展望 24

PART 1

智驾算力运营白皮书编制背景与目的

1.1 智驾算力应用现状

智驾算力正在从研发的"基础设施"演变为决定研发效率与产品竞争力的"战略资源"。

随着端到端模型、VLA模型、世界模型等前沿技术路线的逐步落地，智驾研发已从规则驱动、模块化模型迭代的早期阶段，转向以数据驱动为核心、更大模型的规模化研发为主导的新阶段。算力不再仅仅是支撑模型训练的基础工具，而是直接决定车企智驾研发节奏、模型迭代速度、量产落地时间的核心变量。

然而，算力规模的持续扩张、算力架构的多元异构化、研发多任务并行化，也使得算力管理的复杂度随之大幅攀升。如何高效运营大规模智驾算力、最大化资源利用率、控制研发成本、保障多团队协同效率，已成为整个行业共同关注的关键课题。

本文主要讨论智驾研发过程中使用的算力运营，未涉及车端算力。

1.1.1 智驾算力迈入超大规模集群化时代

当前，智驾模型架构正经历快速演进，从早期模块化感知与决策分离的方案，到BEV、Occupancy Network等统一表征范式的普及，再到端到端大模型及VLA模型的早期量产落地，模型参数量从百万级跃升至百亿乃至千亿级，对训练算力的需求呈跨越式增长。与此同时，智驾研发已形成"数据采集—模型训练—OTA推送—车端运行—数据回传"的闭环体系，数据流转与模型迭代速度不断加快。部分车企与智驾算法供应商已建成数万卡级AI算力集群，支撑百万级智驾车辆的数据回传与模型迭代。智驾算力正式迈入超大规模集群化时代。

1.1.2 智驾算力向多元异构方向演进

当前智驾算力已打破单一GPU架构的建设模式，形成GPU、NPU及其他类型加速芯片搭配使用的异构体系。在训练侧，不同芯片因架构特点、显存带宽、互联拓扑的差异，承担不同的训练任务；在推理侧，不同芯片根据算力、显存、功耗等特征分别适配不同的数据生产任务。目前部分企业已在万卡集群中实现多类型芯片协同训练能力，国产算力在智驾训练中的占比正在快速提升。这一转变的核心价值在于通过多元算力协同，摆脱对单一芯片的依赖，在保障训练性能的同时，实现供应链安全与成本优化。

1.1.3 智驾研发呈多模型、多范式、多任务并行态势

智驾研发不是单一模型迭代，而是多类型模型同步开发、多训练范式并行的复杂工程。模型层面，从参数规模较小的感知模型，到参数规模百亿级乃至千亿级的VLA模型、世界模型，不同模型的算力需求差异极大，前者需要中小规模算力快速迭代验证，后者需要大规模集群支撑长周期训练。训练范式层面，已从预训练拓展到监督微调、强化学习等多种范式，不同范式的算力消耗特征各不相同，对算力调度的灵活性提出了更高要求。任务调度层面，单个研发团队同时有数十乃至上百个训练、推理、仿真验证等不同优先级的任务并行执行，长周期、高资源占用的任务与短时、高频、突发性小任务争抢算力资源，任务优先级高低不一、资源需求波动剧烈，算力资源的分配与调度复杂度显著上升。

1.2 智驾算力运营面临的核心困难与挑战

1.2.1 超大规模算力扩张下的供需错配挑战

受模型演进与数据规模双轮驱动，智驾算力迈入万卡级至十万卡级超大规模阶段，车云数据闭环持续加速研发迭代，算力需求呈爆发式、非线性增长。由于算力需求峰值难以精准预判，算力运营易面临资源供给与研发进度错配的风险，合理配置算力资源、做好前瞻规划与预算管控，成为规模化算力经营的首要命题。

1.2.2 多元异构算力统一度量与运营挑战

伴随算力架构向多元异构融合演进，出于供应链安全与成本优化考量，多类型芯片协同部署、搭配使用逐步落地。当前行业尚未形成统一的算力运营度量规范与指标体系，不同算力资源的能力特征差异显著，传统同构算力环境下的资源计量、配额管理及运营规则难以直接复用，算力运营需破解异构资源价值折算、全域池化调度、跨架构统一经营的难题，避免形成资源孤岛。

1.2.3 多任务并行研发下的资源分配与协同挑战

智驾研发是多种模型、多种训练范式与多任务并行的过程，不同任务的算力需求规模、运行周期、消耗模式差异显著，优先级随研发节点动态变化。长周期大模型训练与高频小模型迭代的资源需求分化明显，多团队并行研发易加剧资源竞争，算力运营需建立科学的任务优先级调度与弹性资源分配机制，平衡核心版本迭代与探索性研发的资源需求，提升整体研发协同效率。

1.2.4 复杂场景下的效能管控与成本平衡挑战

在算力超大规模化、资源异构化、多任务并行化的三重叠加背景下，算力运营的精细化管控难度显著攀升。一方面，海量任务与异构资源深度交织，算力消耗的精准归属、项目级成本分摊难度大幅提升，难以量化不同研发活动的真实算力投入；另一方面，研发试错带来的算力消耗难以事前评估，而资源错配、调度低效产生的无效消耗更具隐蔽性，识别与治理难度加大。同时，保障研发迭代效率、提升资源利用水平、控制整体投入成本三者存在天然的目标张力，实现三者动态平衡的运营难度显著增加。

1.3 智驾算力运营白皮书编制目的

基于当前智驾算力的超大规模部署、多元异构融合、多任务并行研发的现状，以及行业普遍面临的资源规划精准度不足、异构算力治理难度大、多任务协同调度效率低、效能成本管控粗放等核心挑战，本白皮书聚焦智能驾驶研发场景的算力运营体系建设，旨在达成以下目标：

输出体系化方法论，提供企业落地指引

针对智驾研发超大规模算力需求预判、异构算力资源治理、多任务动态调度、效能与成本平衡等核心痛点，结合智驾研发业务特性，提供可落地的运营策略、治理机制与管理方法，为车企及智驾方案商的算力运营团队提供体系化的实践路径与操作参考。

统一度量评价标准，沉淀行业通用语言

针对异构算力场景下行业度量规范缺失的现状，提出通用的算力运营指标体系与价值核算方法，推动形成行业统一的算力运营评价尺度与术语体系，为跨主体沟通、跨企业效能对标奠定标准化基础。

汇聚行业实践经验，共建迭代发展共识

汇集行业各方算力运营实践成果，依托统一的认知框架与度量标准，共同沉淀最佳实践、持续迭代运营方法体系。通过经验互通、方法互鉴，带动全行业算力运营能力共同提升，支撑智能驾驶研发效率稳步提升。

PART 2

算力运营机制

建立明确的算力运营机制，核心目的在于从依赖个人经验与临时决策，转向依靠标准化流程与制度规则。算力运营涉及多团队、多项目对有限资源的争用与协调，若仅依赖口头约定或临时决策，往往面临权责边界模糊、分配规则因人而异、管理动作碎片化等困境。运营机制需要为企业构建一套覆盖完整管理周期的标准化流程与规则体系，涵盖年度资源规划、月度使用分析、季度复盘校准、任务优先级管理与调度，以及周报、月报等运营状态的透明化呈现，确保各环节责任明确、规则统一、执行有据，不因人员变动或项目更迭而中断。运营机制是算力运营从粗放走向精细的重要基础。

以下基于行业最佳实践，梳理算力运营中常见的机制设计，供正在构建或优化自身运营体系的企业参考。

2.1 资源规划管理机制

建立年度配额规划方案，按团队拆分月度费用配额的预算管控机制，实现成本前置约束而非事后核算。

- **预算前置：**年度预算制定先于资源使用，月度配额在年度框架内滚动调整，确保资源消耗始终在预算约束之内。
- **优先级倾斜：**按业务重要程度分配配额，体现资源向高价值业务倾斜的导向。
- **保底与弹性结合：**业务按优先级进行管理，高优先级业务享有配额保底机制，确保核心业务不受配额波动影响；低优先级业务配额可动态调整，保留弹性空间。
- **实时监控：**配额执行情况实时入账，建立分级预警机制，使配额消耗在接近上限时自动触发管控动作，而非依赖人工检查。

2.2 资源申请开通机制

建立"按月审报次月需求"的常态化收集与审批机制，使资源规划前置化、审批规范化，同时为紧急需求预留快速通道。

遵循原则：

- **前置规划：**每月固定周期完成下月需求收集、初审、评审和发布，形成可预期的审批周期，避免临时申请干扰资源分配秩序。
- **分层审批：**运营团队初审把关合规性与可行性，评审会议集体决策分配方案，审批权限与资源影响程度匹配。
- **紧急通道：**对月底评审后突发的紧急需求，建立一事一议的快速审批通道，由指定层级负责人直接审批，事后补录追认，确保紧急业务不被流程阻塞。
- **可追溯：**所有需求申请、审批记录、分配方案统一存档，支持事后审计和复盘分析。

2.3 任务与优先级管理机制

建立统一的任务分级与调度机制，使算力资源在多任务竞争环境下有序分配，高价值任务优先获得保障。

遵循原则：

- **分级保障：**按业务重要程度将任务划分为若干优先级层级，高优先级任务在资源竞争中获得优先保障，体现资源向核心业务倾斜的导向。
- **公平排队：**同优先级任务按提交时间排队，减少资源分配的随意性和人为干预。
- **抢占与回收：**允许高优先级任务在资源不足时优先调用低优先级任务所占用的资源，同时建立闲置资源自动回收机制，避免资源被长期占用而未及时释放。
- **异常感知：**任务运行状态实时监控，异常自动告警并通知责任人，降低异常任务被遗漏的风险。

2.4 资源效率管理机制

建立资源及成本分析报告机制，定期全面反映资源使用情况和成本消耗，为运营决策提供数据支撑。按训练和推理分别分析核心效率指标，通报各团队使用情况，推动效率持续提升。建立基于历史数据和业务规划的分析预测机制，驱动年度和月度计划的动态调整，使资源规划具有前瞻性。

- **全面覆盖：**报告涵盖各团队资源使用量、利用率、费用明细、配额执行情况，确保运营全貌可见。
- **多维对比：**费用数据与月度配额对比标注超支结余，与历史环比同比分析趋势，使成本变化可定位、可解释。
- **指标量化：**建立资源占用率、GPU/NPU利用率、有效训练时间比、多卡线性度等核心效率指标体系，使效率可衡量、可比较。
- **数据驱动：**报告以量化数据为核心，减少主观描述，确保决策依据可量化、可追溯。
- **效率排名通报：**按效率指标排名通报优秀与待改进团队，建立正向激励与整改督促并重的反馈机制。
- **整改闭环：**低效团队需提交整改计划并持续跟踪整改效果，确保发现的问题得到有效整改，并通过复核验证整改成效，形成完整的问题整改闭环。

2.5 周报/月报、季度复盘机制

建立常态化的周报/月报、季度复盘机制，确保运营信息定期同步、问题及时暴露、策略及时调整、决策有据可依。

- **定期发布：**周报和月报按固定周期发布，形成稳定的运营信息周期，使各团队对信息同步有明确预期。
- **分层报告：**周报聚焦效率指标和短期问题，月报覆盖成本分析和趋势研判，季度复盘聚焦阶段性趋势研判与目标达成评估，形成持续改进的长效闭环，报告定位清晰、互为补充。

- **存档可查：**所有报告统一存档，支持按时间、团队、指标维度检索，使运营历史可回溯、可对比。
- **评审闭环：**按训练和推理分别分析核心效率指标，通报各团队使用情况，推动效率持续提升。月报发布后召开运营评审会，讨论问题及改进措施落实情况，确保报告不仅承载信息传递功能，更发挥决策驱动作用。以季度为周期对比核心指标变化，标注改善与劣化趋势，使运营效果可量化评估、可对比。

PART 3

算力运营指标体系

3.1 智驾算力运营指标体系的构建原则与核心场景

科学的算力运营指标体系，既是实时反映算力运营状态的"仪表盘"，也是智驾企业衡量基础设施投入产出、提升模型研发效率、评估业务成果的管理基石。

构建该体系需遵循三项原则：一是目标牵引，明确业务目标、统一指标口径，确保指标可比较、可归因溯源；二是全栈监控，实现从芯片到平台再到业务的全链路覆盖；三是逐层归因，支持结果指标逐级拆解，快速定位问题根因。在场景覆盖上，该体系聚焦智驾模型训练与数据生产两大核心场景，紧扣车企在智驾研发中"提升能力、提高效率、降低成本"三大核心诉求构建。

3.2 智驾算力运营指标体系的分层结构

智驾算力运营覆盖经营、业务、任务、资源等多个管理维度，各维度关注的指标类型、服务对象和决策场景各不相同。若将所有指标平铺呈现，既难以体现指标之间的因果关联，也不利于异常问题的快速定位与归因。因此，有必要按照管理视角的需求进行分层，使指标体系形成清晰的价值传导链与问题追溯链。

本文建议指标体系自顶向下划分为经营层、业务层、任务层和资源层四个层级，形成"价值向上传导、问题向下归因"的因果链：底层资源的健康度决定任务执行效率，任务执行效率决定业务研发速度，业务研发速度最终兑现为经营层面的经营成效；反之，上层指标的异常可沿因果链逐层向下拆解，定位至具体的任务与资源环节。四个层级的定位与关注重点如下：

经营层指标

定位于价值衡量与战略决策，重点关注算力预算决策与扩容论证，回答"算力这笔账是否划算"的问题；

业务层指标

定位于业务产出与效率，重点关注算力投入产出比，回答"算力是否加速了研发迭代与提升了数据生产效率"的问题；

任务层指标

定位于过程效率与稳定性，重点关注任务调度与平台运行的效率和稳定性，回答"任务跑得快不快、稳不稳"的问题；

资源层指标

定位于资源健康与高效利用，重点关注算力资源的利用率与健康度，回答"每一块算力是否物尽其用"的问题。

3.3 智驾算力运营基于四层架构的核心指标参考

以下各层指标实例基于行业实践总结，可作为构建智驾算力运营指标体系的参考。需要说明的是，指标并非标准答案，企业应结合自身业务特点与发展阶段取舍、校准与自定义。

一、经营层核心指标

经营层指标通常是企业根据自身业务特点与经营目标自定义的指标，反映企业核心关注的经营问题，可由下层多个指标及其他业务数据汇聚计算得出。

下表展示当前行业常用的典型经营层指标：

指标	计算方法参考	指标含义
单模型训练成本	（训练算力成本 + 存储成本 + 网络成本 + 平台与人力分摊成本）÷ 可交付评测的模型版本数	衡量交付一个可交付评测的模型版本所需的综合成本，支持按模型、按版本逐级分解核算，是训练侧成本管理的核心指标
闲置资源成本占比	闲置资源成本÷算力总成本×100%	衡量"已购置但未使用"的算力造成的浪费，反映出资源调度或配额管理有问题
算力投入产出效率	有效训练产出 ÷ 算力总投入（交付模型版本数 ÷ 千卡 · 周投入）	衡量千卡每周算力的有效产出，是判断算力投入商业化可持续性、支撑扩容或缩容决策的核心依据
单位有效数据生产成本	数据全链路总成本（采集+处理+挖掘+Clip生产+标注+发布）÷有效Clip总数	衡量数据闭环链路的整体经济性
单Clip标注成本	标注总成本 ÷ 标注Clip数	衡量单条 Clip 样本转化为真值的成本，反映"自动预标注 + 人工校正"标注模式的经济性

二、业务层核心指标

业务层指标通常由企业结合模型类型、研发节奏和业务目标自行定义的，通过统一的归一化口径实现跨模型、跨团队、跨周期的横向对比分析；其分析结果向上汇聚，为经营层投入产出分析提供直接依据。下表展示当前行业常用的典型业务层指标：

指标	计算方法参考	指标含义
千万帧单Epoch千卡训练时长	训练时长÷（训练数据帧数÷10^7）÷ 千卡数（单Epoch口径）	衡量帧数据口径下的标准化训练效率，支撑不同任务、不同集群间的效率对比
千万Clip单Epoch千卡训练时长	训练时长 ÷（训练Clip数 ÷ 10^7）÷ 千卡数（单Epoch口径）	衡量Clip口径下的标准化训练效率，支撑不同任务、不同集群间的效率对比
有效模型版本产出	周期内通过评测并正式交付的模型版本数（版/周或版/月）	衡量算力与研发投入是否转化为可用的模型产出，是业务层向经营层传导价值的核心产出度量
Clip生产效率	周期内生产的有效Clip数 ÷ 周期时长（Clip/天）	衡量可训练原始样本的日产出能力，是数据产能规划与训练数据供给保障的核心依据
有效Clip率	周期内通过有效性校验的Clip数 ÷ 周期内Clip生产总数 × 100%	衡量切分与筛选算法的精准度，有效Clip率影响单位有效Clip成本的高低
有效数据率	周期内通过质检的有效数据量 ÷ 周期内采集数据总量 × 100%	衡量采集车与仿真平台持续产生的原始数据中，真正可用于后续生产的比例，影响单位有效数据的采集成本
数据处理良品率	周期内处理合格的数据量÷周期内处理数据总量 × 100%	衡量数据处理流程的加工质量，良品率影响数据处理成本与处理周期
标注良品率	周期内质检通过的标注Clip数 ÷ 周期内标注Clip总数 × 100%	衡量自动加人工标注的首次通过率，影响单Clip标注成本
自动标注占比	周期内自动标注完成的标注量÷ 周期内标注总量 × 100%	衡量数据标注流程的自动化程度，反映标注环节的降本空间与规模化扩展潜力

三、任务层核心指标

任务层指标具有较强的行业通用性，但不同车企在训练框架、调度平台、数据流水线及业务流程方面存在差异，因此有必要建立统一的指标定义、统计边界和数据采集口径。

从数据获取方式看，任务层指标大致可分为三类：一是可通过调度系统、任务日志和平台元数据直接统计的指标，如任务排队时长、解析成功率、Clip生成成功率等；二是需要跨训练框架、调度系统、Checkpoint记录及故障事件日志等多源数据进行关联计算的指标，如ETTR、断点续训恢复时长；三是需要通过标准化性能测试或可比训练实验获得的指标，如线性度。下表展示当前行业实践中常用的任务层指标：

指标	计算方法参考	指标含义
ETTR（Effective Training Time Ratio，有效训练时间占比）	ETTR = 有效训练时间 ÷ 任务占用资源总时长 × 100%	衡量在一个特定的观测周期，算力节点用于模型前向传播、反向传播及梯度更新的时间占总算力运行时间的比例。
线性度	实际加速比 ÷ 卡数增长倍数 × 100%（加速比 = 单卡训练时长 ÷ N卡训练时长）	衡量多机多卡分布式并行训练时，计算性能随算力规模（卡数）扩展而增长的比例。衡量AI集群硬件网络架构、分布式软件框架及并行算法协同优化能力的核心综合指标。
任务排队时长	任务提交时刻 – 获得资源开始运行时刻的时长，取全部任务的分位	排队时间长说明资源池不足或调度策略不合理，是进行扩容与配额管理的直接依据
断点续训恢复时长	故障发生时刻–任务从最近Checkpoint恢复并继续训练时刻的时长	衡量训练系统从故障中自愈的速度，恢复越慢，故障的算力损耗越大，是ETTR与重训成本的关键控制点
单步训练时长	单步训练时长 = 总耗时 ÷ 完成的训练步数	单步训练时长指模型完成一个完整训练迭代（加载一个 global batch 数据 、前向 、反向 、 梯度同步 、 参数更新）所消耗时间
Checkpoint开销	Checkpoint开销 = Checkpoint写入耗时 ÷ 总训练时长 × 100%	训练过程中，因执行 Checkpoint 保存而导致 GPU 处于等待/阻塞状态的累计时间
检索成功率	成功返回结果的检索请求数 ÷ 检索请求总数 × 100%	衡量检索系统（如向量数据库、并行文件系统、对象存储、倒排索引等）的服务交付质量与系统稳定性
检索时长	单次场景检索请求从发起到返回结果的耗时	衡量的是数据闭环中数据供给环节的速度，场景检索是算法工程师"从一个想法到一批数据"的入口

指标	计算方法参考	指标含义
Clip生成成功率	成功生成的Clip数 ÷ 应切分生成的事件总数 × 100%	衡量数据生产系统将原始海量数据（Raw Data）无遗漏、无损坏地转化为算法训练/测试可直接调用的数据集（Clips）的能力
切分时长	单个事件从触发切分到Clip可用的耗时	衡量Clip生产线的运转速度，切分慢会形成管道积压，直接限制Clip日产能（Clips/天）
一次QA通过率	首次质检即通过的标注量 ÷ 送检标注总量 × 100%	衡量数据标注环节的质量指标，一次通过率低意味着标注流水线在做大量重复劳动，推高单Clip标注成本
Dataset Build成功率	构建成功的Dataset版本数 ÷ 发起构建的总次数 × 100%	衡量Dataset生产与发布环节的执行指标，构建失败会打断训练供料的版本节奏，影响训练任务按时启动

四、资源层核心指标

资源层的 GPU、NPU、节点、存储等原始指标来自以 Prometheus 为核心的标准监控体系，NVIDIA GPU 指标经 DCGM 采集，NPU指标经厂商配套监控工具采集，节点与存储指标分别经节点指标采集器与存储系统指标采集器采集，统一汇入 Prometheus 时序数据库。下表展示当前行业实践中常用的资源层指标：

指标	计算方法参考	指标含义
资源分配率	已发放配额的资源量 ÷ 集群可分配资源总量 × 100%	衡量配额发放策略是否合理，评估算力资源在组织维度的覆盖度
资源占用率	实际运行任务的资源量 ÷ 集群可分配资源量 × 100%	衡量集群资源被训练与数据处理任务实际占用的程度，反映资源池的真实紧张状况
集群AI算力/通用算力利用率	实际消耗算力 ÷ 理论峰值算力 × 100%（AI算力指GPU/NPU等加速卡，通用算力指CPU）	衡量底层算力是否被真正用起来

指标	计算方法参考	指标含义
集群可用率（SLA）	（周期总时长−不可用时长）÷ 周期总时长 × 100%	对外承诺基础设施SLA、评估运维保障能力，集群可用率是训练业务能否持续开展的前提，也是可用性事故的考核依据
显存/HBM利用率	实际使用的显存容量（或HBM带宽）÷峰值容量（或带宽）× 100%	定位训练效率的硬件瓶颈，算力利用率高但显存带宽打满，说明瓶颈在访存而非计算，指导模型切分与批次调优
网络通信占比	分布式通信时间÷单步训练总时长 × 100%	衡量分布式训练的网络开销，通信占比是线性度下降的微观根因，指导网络拓扑优化与并行策略调整
存储可用容量	剩余可用存储空间÷存储总容量 × 100%（同时跟踪绝对量TB/PB）	容量不足会导致上传失败与数据丢失，是存储采购与生命周期治理（冷热分层、到期清理）的依据
CPU/GPU利用率（数据处理）	处理任务实际消耗的算力÷分配的处理算力 × 100%	数据处理环节的算力资源指标，利用率长期偏低说明管道并行度不足或任务调度不均，推高单TB处理成本
存储吞吐	处理管道单位时间对存储的读写数据量（GB/s）	衡量处理管道的IO供给能力，吞吐不足时算力再多也只能等数据，是处理周期长短的底层约束
检索算力利用率	场景检索实际消耗的算力÷分配的检索算力 × 100%	场景挖掘环节的算力指标，利用率接近上限时检索时长增加，可指导检索集群扩缩容
读写吞吐（Clip生产）	Clip生产管道单位时间读写数据量（GB/s）	Clip生产环节的存储资源指标，读写吞吐决定切分时长，是Clip日产能（Clips/天）的底层天花板
生成算力利用率	Clip生成任务实际消耗的算力÷分配的生成算力 × 100%	衡量Clip产线的算力投入产出匹配度。利用率不足说明算力虚配，利用率超过一定阈值预示产能瓶颈
自动标注算力利用率	自动标注任务实际消耗的算力÷分配的自动标注算力 × 100%	衡量自动标注产线的算力饱和度，是提升自动标注占比的产能前提，利用率超过一定阈值意味着自动化扩量必须先扩算力
缓存吞吐	缓存层单位时间可支撑的读写数据量（GB/s）	衡量数据集发布后对训练侧的供数能力。缓存吞吐不足会使GPU空闲，影响ETTR与算力利用率

PART 4

算力运营平台建设

4.1 平台参考架构

智驾算力运营平台应以支撑数据驱动的运营闭环为核心。本文建议运营平台应围绕数据底座、运营中心、优化中心等核心模块进行建设，形成覆盖数据汇聚、运营管理与持续优化的算力运营体系。

数据底座

算力运营平台的基础设施层，支持多区域多元异构算力接入，负责算力指标数据的建模、采集、处理和持久化存储，为上层运营中心提供统一的数据支撑。

运营中心

算力运营平台的核心业务层，基于数据底座提供的指标数据，实现算力运营的可视化、预警和报表分析能力。

优化中心

算力运营平台的业务闭环层，基于运营中心的分析结果，明确优化方向，支撑优化动作的执行，持续提升算力资源的使用效率。

参考架构图如下：


优化中心
算力资源管理
资源扩容
资源缩容
资源弹性伸缩
算力资源优化
多数据集
智能调度
卡虚拟化
线性度优化
显存预热
… …
运营中心
指标可视化看板
经营类指标看板
业务类指标看板
任务类指标看板
资源类指标看板
智能巡检预警
算力指标数据分析
算力指标异常预警
算力指标定时巡检
算力巡检数据采集引擎
算力运营报表
算力运营报表
算力运营数据分析
算力指标数据清洗
算力运营数据采集引擎
数据底座
数据存储
时序数据库
多级降采样存储
冷数据存储备份
数据处理
指标数据订阅
指标数据清洗
指标数据聚合
指标数据入湖
数据采集
应用运维管理
云日志服务
数据建模
资源池
资源队列
节点
智算卡
任务
用户
区域A－GPU集群
区域A－NPU集群
区域B－NPU集群
… …

4.2 核心功能模块说明

数据底座

- **数据建模：**对算力的全对象标准化建模，包含算力集群、资源池、资源队列、服务器/节点、智算卡、任务、用户等，为数据底层提供标准的数据统计规则。
- **数据采集：**通过应用运维管理、日志服务采集资源、任务等维度的指标数据。
- **数据处理：**通过消息队列完成指标的实时订阅与分发；通过分布式流式处理框架对原始指标数据进行清洗和过滤，再对指标数据进行预聚合计算，生产统计指标，最后将处理后的指标数据写入到时序数据库，从而完成指标数据入湖。
- **数据存储：**指标数据写入到时序数据库，支持超长周期的指标查询和秒级查询能力；提供指标数据的标准查询接口，对外提供统一的数据底座。

运营中心层

- **指标可视化看板：**定义核心指标，并根据企业需要，构建经营、业务、任务、资源等指标可视化看板，可视情况进行更多的二级看板设计。
- **智能巡检预警中心：**定期巡检算力核心指标，依托Agent智能体提供巡检报告，并结合指标阈值，冒泡异常情况，发起预警通知与及时响应。
- **算力运营报表中心：**基于统一的数据源，进行算力指标数据的采集与分析，提供按日、周、月、季度及年度的算力运营的报表。

优化中心层

- **算力资源管理：**根据算力核心指标来决策资源扩容或缩容时机和规模，支撑算力资源高效使用。
- **算力资源优化：**根据算力运营资源优化的最佳实践提供优化方向指导；以推理与训练为例，比如数据生产调优包括但不限于多数据集并行、AI卡虚拟化、显存预热、容器智能调度、线性度优化等；模型训练调优包括但不限于CP/SP并行策略、重计算关闭、CPU绑核等。通过多种手段组合调优，有效支持算力/显存利用率提升。

PART 5

典型算力运营优化策略及案例

面对算力规模持续扩张、资源架构多元异构、研发任务复杂并行的多重挑战，如何将算力从 "可用" 推向 "好用、用好"，已成为智驾企业运营体系建设的核心命题。本章围绕典型算力运营优化策略展开系统论述，从资源调度与利用率提升、异构算力统一纳管、成本精细核算与控制、多团队协同保障等关键维度，梳理业界行之有效的方法论与实施路径；同时结合代表性企业的落地实践案例，剖析策略实施过程中的关键动作、量化成效与经验教训，助力企业在算力投入与研发效率之间实现最优平衡，为大规模智驾研发的持续提速提供坚实支撑。

5.1 典型算力优化目标及策略

核心优化目标	主要优化策略
算力需求精准化、任务规范化、接入服务化、保障分级化	① 需求标准化+智能准入校验（超配拦截/配额校验） ② 任务全生命周期精细管理（五阶段闭环/闲置强终止） ③ 业务分级分类+QoS资源保障（核心/重要/普通三级） ④ 算力服务化封装（商城化、一键交付） ⑤ 需求预测+前置容量规划（削峰填谷） ⑥ 闲置算力自动回收+业务考核机制
算力利用率提升、资源碎片率降低、任务排队时长减少	① 全域算力池化+异构资源统一纳管 ② 智能调度算法引擎（双层调度+评分机制） ③ 动态负载均衡+算力碎片整合 ④ 潮汐算力调度+错峰运行（削峰填谷） ⑤ 跨域算力路由+任务热迁移 ⑥ 业务分级QoS+抢占式调度 ⑦ 调度自愈+异常容错

核心优化目标	主要优化策略
GPU/NPU利用率提升	①Host Bound根治优化（DataLoader/GPU化/零拷贝） ②通信掩盖优化（通信计算重叠/3D并行/梯度压缩） ③算子优化（算子融合/硬件亲和算子/编译优化） ④模型参数配置优化（Batch/精度/并行策略/超参） ⑤可靠性优化（断点续训/分布式容错/数据可靠性） ⑥算法优化（模型轻量化/稀疏化/高效训练算法） ⑦虚拟化+显存预热
硬件可用性提升	GPU/NPU算力集群、CPU服务器、AI加速卡、交换机、存储阵列等硬件资源池的纳管、故障监控、能耗管控

5.2 算力运营优化案例

案例1：通过优化模型性能，提升数据生产效率

背景：在智驾数据生产中3D感知的Occ-Former模型推理中，推理效率是直接影响算力成本的主要因素，在对Occ-Former执行推理的整个过程中，分析发现GridSampler2dV2Grad算子耗时占整体推理耗时5.97%，成为主要性能瓶颈。现有实现存在大量重复计算、串行标量运算、内存访问碎片化、计算与数据加载不同步等问题，无法充分发挥NPU硬件算力，增大了端到端推理时延。

优化策略流程：采用“问题定位 -> 中间结果复用策略 -> 计算流程重构 -> 精度与性能验证”的递进式优化流程。

步骤1：问题定位

经分析，该算子反向计算存在三大问题：

- **计算逻辑重复：**多个梯度计算路径（gix、giy、giz）均依赖相同中间量gOutLocalTensor * inputXLocalTensor * coorValue，但未进行复用；
- **冗余运算频繁：**相同表达式被多次重复计算，增加算子执行时间；
- **未充分利用硬件特性：**缺乏对昇腾NPU内存访问与向量化计算的深度优化。

步骤2：中间结果复用与逻辑重构

提出“一次计算，多次复用”优化策略：

- **中间量提取与复用：**gix/giy/giz三条梯度计算路径均依赖inputXLocalTensor = gOut-LocalTensor*inputXLocalTensor*coorValue，原实现分别执行造成冗余。优化为先计算一次该中间量并存储，后续所有依赖计算直接复用，避免重复乘法运算；
- **计算流程重构：**先通过Mul（gOut×inputX）+Muls（×coorValue）计算核心中间量，再复用该中间量分别计算gix（×yVal→×zVal）、giz（×xVal）、giy（×xVal→×zVal），减少3次Mul+3次Muls运算，降低计算密度，减少内存访问频次，提升数据局部性；
- **辅助函数优化：**引入CeilDiv（向上取整分块数量）与CeilAlign（数据对齐至硬件块大小）内联函数，确保数据分块合理、规避边界处理开销，提升内存访问效率。

流程主线：Profiling定位瓶颈(5.6s)->分析冗余计算->提取中间量一次计算多次复用->计算流程重构->对齐优化 -> 精度性能验证(约20%)

③结果：在智驾业务标准Shape下，GridSampler2dV2Grad算子性能提升约40%，精度保持业务可用、无精度损失。

④影响指标

指标名称	变化
单算子时延	显著下降，核心瓶颈算子耗时压缩
模型端到端推理时延	降低，提升智驾推理实时响应能力
NPU 算力利用率	消除冗余浮点运算，有效算力占比提升
HBM带宽利用率	访存合并+搬运计算流水，减少stall，带宽利用率提升
缓存命中率	Tile分块、索引查表优化，缓存命中率提升

案例2：通过多种优化方法，提升集群利用率

背景：在算力运营体系中，算力利用率是衡量算力资源实际有效发挥程度的核心运营指标，智驾行业此前长期以芯片纸面峰值TOPS作为算力实力评判标准，大量算力在软硬件适配、底层调度环节被无效消耗，算力利用率成为下半场算力运营的核心考核维度，直接决定智驾大模型训练、路测数据闭环迭代的效率与算力成本管控水平。

结合智驾业务真实的路测数据闭环训练流程验证，Qwen3-VL-8B模型在算力集群上开展多轮智驾场景感知能力微调时，全周期平均NPU资源利用率仅65%，部分分布式训练节点在梯度跨卡同步环节，最低利用率仅40%。当前利用率水平无法支撑智驾模型快速迭代的算力运营效率要求，大量算力在数据等待阶段处于空转状态，优化潜力充足。

优化策略流程：

步骤1：开启NPU亲和绑核

修改环境变量CPU_AFFINITY_CONF=2,npu_affine:1，完成 CPU NPU 亲和绑定，减少调度抖动，利用率由65%提升至67%，提升2%

步骤2：调优微批micro_batch_size（MBS）

在绑核基础上调大微批，提升单轮计算的数据负载：

- micro_batch_size=2：利用率提升至70%；
- micro_batch_size=4：利用率无进一步上涨，维持70%。

步骤3：调高重计算层数

将recompute_num_layers从1调整至18，以计算换显存，增加模型计算量。

步骤4：优化dataloader数据加载并行参数

调整dataloader_num_workers=32、dataset_num_proc=16，提升数据预处理并行能力。

步骤5：开启序列Packing

开启packing=true，设置packing_num_proc=8，消除输入padding无效填充，提升有效数据占比，NPU利用率可达80%；

步骤6：切换DeepSpeed分布式训练框架

原Megatron框架张量并行带来频繁节点通信；切换DeepSpeed框架，优化梯度通信逻辑，降低通信开销，资源利用率稳定75%以上；

步骤7：Profiling定位，删除显存卸载类参数

通过Profiling性能分析定位，删除optimizer_cpu_offload、use_precision_aware_optimizer、optimizer_offload_fraction等显存卸载参数，减少NPU->CPU之间频繁数据搬运。

③ 结果：综合对比全部方案，筛选出提升微批大小、开启Packing、切换DeepSpeed框架三类低成本高收益方案落地。NPU集群平均资源利用率由65%提升至≥75%，满足算力运营要求；利用率最低值由40%抬升至75%，消除利用率低谷。

④ 影响指标

指标名称	说明
跨节点通信利用率	框架选型影响通信耗时，过高拉低有效利用率
数据预处理吞吐	dataloader线程数影响数据供给（本案例非瓶颈）
训练有效样本占比	Packing消除padding，提升有效数据占比
CPU-NPU搬运开销	卸载参数触发频繁搬移造成算力空转

PART 6

算力运营成熟度模型

6.1 构建算力运营成熟度模型的意义

前文已基于行业实践系统梳理了智驾算力运营的核心机制、指标体系与平台参考架构。在企业落地算力运营体系的过程中，如何客观评估自身运营水平、明确能力建设优先级，是行业普遍面临的现实问题。受算力规模、团队能力、业务发展阶段等差异影响，企业算力运营能力建设无法一蹴而就，也难以套用统一模式。因此，构建一套分层分级、循序渐进的成熟度评估框架，是帮助企业评估现状、精准定位、规划演进路径的必要基础。

引入智驾算力运营成熟度模型，对行业发展具有三重核心价值：

提供自我评估的客观参照标尺

当前多数车企对自身算力运营水平缺乏系统性、量化的客观认知，难以精准识别能力短板与改进方向。成熟度模型构建了标准化的评估体系，支撑企业快速定位自身发展阶段，明确能力缺口。

锚定能力演进的清晰建设路径

成熟度等级并非单纯的分级标签，而是一套阶梯式的能力建设指引。企业可基于当前所处等级，明确下一阶段的目标状态与核心建设任务，避免盲目投入或跨越式建设导致的资源浪费。

推动行业形成统一的运营共识

当前智驾算力运营领域缺乏通用的评估标准与术语体系，不同主体对运营能力的评判尺度差异较大。成熟度模型提供了统一的语言框架，可提升企业间、企业与服务商间的沟通效率，为后续行业标准沉淀与最佳实践推广奠定基础。

6.2 算力运营成熟度等级定义

本白皮书将智驾算力运营能力划分为逐级递进的五个成熟度等级，从监控能力、度量体系、调度方式、优化手段、智能化程度五个维度进行分级评估。各等级逐级递进，后一级别以前一级别的能力为基础，体现运营能力从被动到主动、从人工到自动、从局部到全局的演进路径。各等级的核心特征如下表所示：

成熟度等级	名称	特征描述
L1	初始级	缺乏系统化的监控与运营手段，算力使用状态不透明。故障发现依赖人工巡检，问题响应以被动处置为主。资源分配依靠人工协调，无统一调度平台，团队间资源争抢与闲置错配现象并存。算力成本核算方式粗放，无法实现任务级、团队级的精准归属。
L2	可观测级	建成基础监控体系与指标采集能力，可对集群硬件状态、任务运行状态开展实时监控与可视化展示。运维团队可实现算力使用状态的可视化感知，但决策判断仍依赖人工经验。初步具备故障告警能力，告警规则以阈值触发为主，存在一定的误报与漏报。
L3	可度量级	建成覆盖硬件利用率、训练效率、资源周转效率、成本效能等维度的完整指标体系，形成跨团队、跨任务统一的度量口径。可量化评估不同模型、不同训练范式、不同团队的算力使用效率，为资源分配决策与成本优化提供数据支撑。具备多维度报表与归属分析能力，可有效支撑算力投向、投入效能等核心问题的研判。
L4	可优化级	在可度量能力基础上，具备自动化调度与主动优化能力。系统可基于指标数据自动识别效率瓶颈，并通过智能调度、弹性伸缩、任务编排等手段主动优化资源分配。具备故障快速定位与自动恢复能力，训练中断恢复时间从小时级缩短至分钟级，实现从"人工排查问题"向"系统主动发现并驱动问题解决"的转变。
L5	智能化级	基于AI与大数据技术构建智能化运营闭环。系统可基于历史运营数据开展容量预测与自动规划，在业务需求落地前完成资源预分配；具备故障预测与预防性维护能力，可在硬件故障发生前主动规避风险；实现资源自优化与自演进，可根据业务负载变化自动调整调度策略，持续逼近全局最优状态。运营团队角色由执行操作向策略制定与全局管控转型。

6.3 算力运营成熟度应用建议

企业在应用本成熟度模型评估自身算力运营能力现状、识别能力短板、规划演进路径时，建议遵循以下原则：

如实评估，避免跳跃式建设。企业应先夯实监控与度量基础，再逐步向自动化、智能化演进，确保每一阶段的能力扎实落地。

分阶段设定目标，逐步迭代。建议企业以年度为单位设定成熟度提升目标，每个阶段聚焦2~3个关键能力建设，避免一次性铺开过多建设任务导致资源分散。

结合业务节奏动态调整。成熟度建设应与企业的智驾研发节奏相匹配，在模型快速迭代期优先提升调度灵活性，在规模化量产期优先提升成本管控与稳定性保障能力。

PART 7

行业倡议与展望

当前智能驾驶产业已进入大模型驱动、数据闭环迭代、超大规模算力集群支撑的发展阶段。算力已成为车企智驾研发迭代、技术创新与产品落地的核心战略资源。伴随算力规模持续扩张、研发任务日趋复杂，行业普遍面临运营标准不统一、资源利用效率偏低、产业经验互通不足、大规模任务交付稳定性有待提升等共性问题。为破解行业痛点、构建体系化算力运营能力、推动产业协同高质量发展，本白皮书联合多家车企与算力提供商，提出以下四点行业倡议。

统一行业度量标准，构建标准化算力运营体系

当前智驾算力应用场景多元、技术路线并行，行业尚未形成通用的算力度量与运营评价规范，不同算力资源的效能、价值、利用率无法横向对标，企业间管理口径不一致，制约了行业整体精细化运营水平的提升。

倡议全行业共建智驾研发算力运营统一标准体系，统一算力资源计量口径、利用率统计规则、任务效能评价维度、项目成本分摊方法，形成行业通用、可对标、可追溯的算力运营评价语言。打破不同架构、不同企业间的标准壁垒，为产业效能对标、经验复用、协同发展奠定标准化基础。

深化精细化算力治理，提升全域资源利用效能

随着智驾算力集群迈入超大规模阶段，多模型、多任务、多优先级研发并行常态化，传统粗放式资源管理模式易引发资源错配、任务排队、算力闲置、无效消耗偏高等问题，造成研发算力投入与产出不匹配。

倡议行业主体全面推进全链路精细化算力运营治理，建立"事前需求预判、事中动态调度、事后效能复盘"的闭环运营机制。结合智驾研发节奏动态优化自建算力与弹性算力配比，针对不同类型研发任务实施差异化资源分配与优先级调度，常态化治理低效算力消耗，实现研发效率、资源利用率与算力成本的最优平衡。

共建产业协同机制，推动最佳实践共享复用

当前行业算力运营能力发展不均衡，部分企业已形成成熟的规模化运营体系与方法论，不少企业仍存在运营体系不完善、调度能力不足、成本管控薄弱等问题，行业一定程度上存在重复建设与重复试错的情况。

倡议行业搭建常态化算力运营协同共享机制，依托产业联盟、行业交流平台，开放算力调度体系、资源治理方案、效能优化经验、运营平台建设等成熟实践。鼓励车企与产业链上下游互通运营经验、共享落地案例，降低全行业体系建设与试错成本，推动优质运营能力规模化复用，整体提升国内智驾算力运营成熟度。

完善全流程运营管理，保障研发高效稳定交付

智驾研发算力承载模型训练、数据生产、仿真推理等核心业务，随着算力规模持续扩大、任务并发度不断提升，算力调度的有序性、运行的稳定性、交付的可靠性直接影响研发迭代效率。行业部分企业仍存在管控粒度较粗、调度规则不清晰、异常响应不及时等问题，可能影响研发任务推进节奏与交付质量。

倡议全行业建立规范化、全流程的算力运营管理体系，完善资源配额管理、任务调度规范、运行实时监控、异常快速响应等运营机制。强化算力资源全生命周期管控，保障超大规模集群、高并发任务场景下算力资源有序调度、稳定运行、可靠交付，为智驾模型持续迭代与研发业务高效推进提供坚实支撑。

展望

智驾研发业务的算力运营是支撑智能驾驶技术持续迭代、产业规模化升级的关键底座能力。



未来行业需持续推进算力运营标准化、精细化、协同化、体系化建设，以统一标准规范产业发展、以精细运营释放算力价值、以生态协同降低行业成本、以体系运营保障研发效能，持续加速智驾技术创新与量产落地，共同推动中国智能驾驶产业高质量发展。

白皮书作者及修订情况

版本： V1.0

主要作者（作者名单按姓名拼音首字母排序）

陈　雷　陈　涛　陈晓聪　方宇行　何　叶　黄　睿
刘　勇　马伟峰　潘文清　彭永成　孙　思　孙　卫
王桂川　汪　航　武帅东　向　阳　谢文敏　熊　建
熊林坤　薛　炜　张超鹏　张　宏　赵　刚　周凯旋
朱　军

修订情况： 初始版本

版权所有©华为技术有限公司2026。保留一切权利。
非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。

商标声明
和其他华为商标均为华为技术有限公司的商标或者注册商标。
在本白皮书以及白皮书描述的产品中，出现的其他商标、产品名称、服务名称以及公司名称，由其各自的所有人拥有。

免责声明
本文档可能含有预测信息，包括但不限于有关未来的市场、运营、产品、技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺。华为可能不经通知修改上述信息，恕不另行通知。