MemoryLake

客户版 · CLIENT

MEMORYLAKE COMPUTE

产品说明书

客户版

以记忆为中心的智能计算平台 —— 以标准计算单元 MCU 为交付粒度，
为企业提供高性能 AI 计算资源的按需使用与统一管理。

版本	日期	出品
v2.1	2026-06-30	质变科技

一、产品定义

MemoryLake Compute 计算平台 是面向企业的、以记忆为中心的智能计算平台服务，以标准化计算单元 MCU（MemoryLake Compute Unit） 为交付粒度，为客户提供高性能 AI 计算资源的按需使用与统一管理能力。

客户购买的是标准化、可度量、有 SLA 保障的计算能力，而不是硬件设备。MCU 之于算力，就像"度（kWh）"之于电力——您按计算能力采购，平台负责底层资源的供给、运维与保障。

1.1 产品边界

MemoryLake Compute 是	不是
平台化、按计算能力售卖的服务	GPU 服务器 / 裸机租赁
以 MCU 标准单元交付与计量	按硬件台数 / 卡数交付
含管理软件 + 运维 + SLA 的完整服务	仅提供硬件的单一交付
客户通过控制台 / API 统一使用与管理算力	客户自行管理底层设备与运维

1.2 价值主张

- 免基建：无需采购硬件、建设机房、组建运维团队，开通即用。
- 可度量：算力以 MCU 标准化度量，用了多少、剩多少、跑了什么，一目了然。
- 有保障：平台级 SLA 承诺，节点故障自动迁移，不中断业务。
- 易管理：统一控制台管理资源、用量与账单，按需弹性伸缩。
- 可扩展：在资源池额度内弹性伸缩，峰值需求可按量补充。

1.3 目标客户与典型场景

面向有 AI 计算需求、但不想或不能自建 GPU 基础设施的企业：

- 大模型训练与微调：预训练、继续预训练、SFT / LoRA 微调。
- AI 推理服务部署：在线推理、批量推理、多模型托管。
- 大规模数据处理与分析：向量化、embedding 生成、特征工程、离线计算。

二、MCU 标准计算单元规格

2.1 MCU 是什么

> MCU（MemoryLake Compute Unit） 是 MemoryLake Compute 平台定义的标准 AI 计算能力单位，是一个固定配比的原子资源包。每个 MCU 包含 AI 算力、计算显存、系统内存、本地高速存储与高速网络接入，平台保证每个 MCU 的实际处理性能达到下表承诺。

2.2 每 MCU 交付能力（承诺规格）

能力维度	每 MCU 保证	说明
AI 计算能力	⩾ 125 TFLOPS（FP16/BF16 Tensor）※	面向训练与推理的张量算力
计算显存	⩾ 21 GB	高带宽 GPU 显存
系统内存	⩾ 48 GB	DDR5 高频内存
本地数据存储	⩾ 360 GB NVMe	高速本地盘，任务级数据缓存
vCPU	⩾ 8 vCPU	用于数据加载、预处理与编排
高速计算网络	高速互联接入（MCU 间紧耦合通信）	支持分布式训练的高带宽低延迟互联

※ AI 算力承诺值以平台上线实测基准为准，合同 / SLA 以平台公布的当期承诺指标为准。

> 规格只描述计算能力指标（TFLOPS、GB、网络能力），不绑定任何 GPU 型号、卡数或服务器配置。 平台保留在不降低承诺规格的前提下迭代底层实现的权利。

2.3 规格选择：数量 + 地域 + 调度域

MCU 是统一原子单元，客户按以下维度选型：

维度	选项	适用
数量	任意整数 MCU（建议按 4 的倍数采购）	与负载规模匹配
地域	多地域可选，按各地域库存与价格选择	就近部署、合规与时延需求

维度	选项	适用
调度域	普通调度 / 紧耦合计算域	紧耦合计算域内的 MCU 共享高速互联，分布式训练必选；推理 / 数据处理可用普通调度

推荐场景套餐（选型建议，本质是 MCU 数量 + 调度域组合）

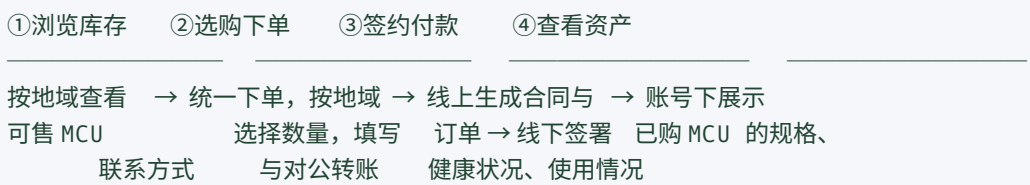
场景套餐	建议 MCU	调度域	适用负载
推理 / 开发套餐	4 MCU	普通调度	在线推理、模型调试、轻量任务
微调套餐	8 MCU	紧耦合计算域	LoRA / SFT 微调、中等规模推理集群
训练套餐	32 MCU（单一紧耦合计算域）	紧耦合计算域	分布式训练、大模型微调，保证域内高速互联
集群套餐	64 MCU 起，按需扩展	多紧耦合计算域	大规模训练 / 大规模推理平台

> **紧耦合计算域**：平台保证同一计算域内的 MCU 通过高带宽、低延迟网络互联，适合需要频繁参数同步的分布式训练。客户无需关心底层拓扑，平台按调度域承诺互联能力。

三、平台功能与使用流程

> 当前阶段聚焦一条清晰主线：**浏览库存 → 选购 MCU → 下单签约付款 → 查看算力资产**。客户全程通过 Web 控制台自助完成，无需关心底层资源调度与运维。

3.1 端到端流程总览



3.2 库存浏览

功能	说明
可售库存总览	查看平台当前可售 MCU 总量与实时可用情况
按地域查看	库存按地域分别展示可售 MCU、单价与负载水位；当前开放 华东 · 上海、华北 · 北京、华南 · 深圳、西部 · 成都 等多个地域
规格说明	每 MCU 的标准能力规格（见 §2.2）随库存一并展示，便于选型

3.3 选购与下单

步骤	说明
选择地域与数量	统一下单：在一张订单中按地域分别选择采购的 MCU 数量（建议按 4 的倍数）。不同地域价格可不同（当前各地域价格一致）
（可选）调度域	分布式训练可选紧耦合计算域，保证 MCU 间高速互联（见 §2.3）
填写联系方式	填写采购方名称、联系人、联系电话 / 邮箱等签约与开通所需信息
提交下单	提交后平台自动生成订单信息与服务合同（含分地域明细与总额）

3.4 合同签署与支付

> 当前阶段采用线下签约 + 对公转账模式。

步骤	说明
合同生成	系统按所选地域、MCU 数量、售卖形态自动生成服务合同与订单总额
线下签署	双方完成服务合同的线下签署
对公转账	客户按订单金额完成对公银行转账（收款主体：杭州质变科技有限公司）
平台确认开通	平台收款并核验后，将订单状态置为"已完成"，并将对应 MCU 资源交付至客户账号；客户账号看板随即生效

3.5 我的算力资产（购后可视化看板）

签约付款完成后，客户账号下即展示已购算力资产的可视化指标：

看板	展示内容
已购 MCU 概览	已购 MCU 数量、规格、所在地域、服务期限、生效 / 到期状态
资源利用率	平均利用率、活跃 / 空闲 MCU、运行中任务、本月已用 MCU · 时
用量趋势	近 7 日 MCU · 时用量趋势
健康状况	以 MCU 点阵呈现，每个方块代表 1 个 MCU，按状态着色（正常 / 告警 / 迁移中），整体健康一目了然
分地域明细	各地域已购 MCU、利用率、健康状态与月度服务费
账单与用量	月度账单、MCU 消耗明细，支持导出

> 节点上下线、负载均衡、故障迁移、硬件维护、容量规划等均由平台侧自动完成，客户无感知、无需干预。

3.6 接入方式

方式	适用场景	状态
Web 控制台	浏览库存、下单、查看算力资产（主入口）	已上线
API / SDK	自动化集成、程序化资源管理	规划中

方式	适用场景	状态
CLI 工具	开发者本地操作	规划中

四、服务等级协议（SLA）

> 以下为平台承诺框架，具体数值以签约 SLA 条款为准。

指标	承诺
平台月度可用性	⩾ 99.9%
单 MCU 性能达标率	⩾ 99%（按承诺规格 §2.2 衡量）
节点故障任务迁移	自动迁移，业务无需人工干预
工单响应	严重故障 ⩽ 30 分钟响应，7×24 支持
数据隔离	租户间计算 / 存储 / 网络强隔离

五、计量与计费

5.1 计量单位：MCU

客户按 MCU 购买计算能力。计量基于 MCU 分配量 与 MCU 使用时长。

5.2 工作负载消耗系数

不同负载对资源的消耗强度不同，通过消耗系数调节实际计量：

工作负载类型	消耗系数	计量逻辑
模型训练（Training）	1.5×	1 MCU 跑 1 小时训练 = 计 1.5 MCU · 时
模型推理（Inference）	1.0×	基准负载
开发调试（Development）	0.8×	交互式、间歇占用
数据处理（Data Processing）	0.6×	CPU 密集、算力占用低

> **包月 / 包年客户**：月度费用固定，消耗系数仅影响资源池内的**使用效率**（即不同负载折算占用多少 MCU · 时），不额外计费。鼓励客户用轻量任务填充闲时资源、提升利用率。

5.3 售卖形态与定价

形态	计费单位	MCU 单价	适用
资源池包年（主推）	MCU / 月（年付）	¥1,080 / MCU / 月	长期稳定需求，锁定价格，享年付权益
资源池包月	MCU / 月	按当期报价	持续计算需求
按量计费	MCU · 时	按当期报价	短期项目、弹性峰值

价格梯度：按量 > 包月 > 包年（体现承诺折扣）。

> **分地域定价**：不同地域的 MCU 单价可不同；**当前各地域价格统一为上表值**。下单时按所选地域分别计价，订单总额为各地域明细之和。

5.4 账单与发票

```
月度账单
├── 账单周期：YYYY-MM-01 ~ YYYY-MM-DD
├── 服务项目：MemoryLake Compute 智能计算平台服务
├── 资源池规格：N MCU（按地域分项）
├── 月度服务费：¥Σ(地域 MCU × 地域单价)
└── 发票品名：信息技术服务费（税率 6%）
```

六、采购指引

完整购买流程见 §3.1～§3.5。选型要点：

1. 评估负载规模 → 估算所需 MCU 数量（参考 §2.3 场景套餐）。
2. 确定地域与调度域 → 就近选择库存充足的地域；分布式训练选紧耦合计算域。
3. 选择售卖形态 → 长期稳定首选包年，弹性峰值用按量补充。
4. 下单 → 签约 → 付款 → 平台自动生成订单与合同，线下签署并完成对公转账后，平台确认开通。
5. 查看算力资产 → 账号下查看已购 MCU 的健康状况与使用情况。

合同表述示例

```
服务项目：MemoryLake Compute 智能计算平台服务
资源池规格：N MCU（AI 总算力 ≥ N × 125 TFLOPS，FP16/BF16）
月度服务费：¥(N × 1,080) 元/月（多地域时按地域分项合计）
服务期限：12 个月
发票品名：信息技术服务费（税率 6%）
```

本文档面向客户与商务场景，仅含客户可见信息。底层资源配置、单元化换算与成本 / 定价推导等平台机密信息，详见《MemoryLake_Compute_产品说明书_机密版》（仅限内部，不得向客户或渠道披露）。