火山引擎

探索能源领域的数字化增长引擎

火山引擎
能源行业高级解决方案总监
孙博文

01 公司介绍

伴随数字化趋势发展诞生的火山引擎

火山引擎

PC 移动

个性化推荐

3G 4G

视频化

AI

大模型

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

2012.7 上线今日头条

2016.6 上线西瓜视频

2016.9 上线抖音

2017.8 上线懂车帝

2019.4 上线飞书

2019.5 上线剪映

2021.6 推出火山引擎

2021.8 PICO加入

2023.6 推出火山方舟

火山引擎能力矩阵

火山引擎

火山引擎提供云基础、视频与内容分发、数智平台VeDI、 人工智能、开发与运维等服务，帮助企业在数字化升级中实现持续增长。

公司资质

火山引擎已经通过多项国际 ISO认证

网络安全等级保护三级

可信云认证

CSA Star认证

SOC 2 审计

隐私信息管理认证

质量管理体系认证

云服务信息安全管理体系认证

公有云中个人信息保护认证

业务连续性管理体系认证

个人身份信息保护实践认证

信息安全管理体系认证

IT服务管理标准认证

02 今年热点—大模型

大模型促进AI逐渐通用化

火山引擎

通用AI技术应用

搜索引擎和APP理解你的语言 ← 自然语言处理

语义理解
语音识别
} 语音助手及APP里的“新玩具”

机器读图 ← 图像处理

特征识别 → 对海量的UGC内容自动标注

人脸识别 → ToB：公共安全
ToC：刷脸支付

大数据系统根据设计进行分类等 ← 数据处理

大数据分析 → 已成为大量行业的基础

千行百业就有千百个模型

每个模型重新训练，建设成本高

市场分散

预训练模型，将每个模型的部分工作前置

减少重复工作量

行业AI技术应用

机器视觉
+
知识图谱
+
文字生成

机器视觉
+
专家系统
+
(排产优化)

每个模型由对应的人使用

对人员要求高，效率低

未改变行业形态

让“大模型”来使用“小模型”

大模型在能源领域功能性场景

火山引擎

01

支撑企业管理

- **数据洞察：**充分利用数据中台沉淀，以自然语言控制灵活的数据洞察。
- **内部知识搜索与问答：**面向能源运检等大量规程规范，实现快速查询
- **报告辅助写作：**根据样例总结，或自动转换不同类型报告文风。

02

智能化企业（2024年底）

- 大模型与管理系统深度耦合，实现工作汇报半自动化，AI参与到可量化工作日常监督、汇报总结等工作中
- 大模型与生产系统出现轻度耦合，在对于实时数据智能数据洞察基础上，增加决策智能。

03

生产智能化（2025年）

- 通过预训练模型，进行故障预警、性能优化等AI的研发，减少不同机组、不同设备间差异造成的成本上升，在生产领域应用大规模推广。
- 以求解器等决策智能为核心，以LLM为交互及工作流引擎，准实时形成适应高新能源装机比例及柔性潮流控制等新型手段，自动化运行阻塞计算等工具，快速生成电网调度方式策略。

从对话到功能的大模型工程化

火山引擎

用户输入：我要准备一篇关于AI agent的科普演讲，时长3分钟，请帮我列出大纲。

Chat类大模型

直接给出答案：

一、定义和背景
什么是 AI agent
为什么研究 AI agent
二、AI agent 的分类
根据功能分类
根据应用场景分类
三、核心技术
机器学习
自然语言处理
计算机视觉
四、AI agent 的应用
智能客服
智能推荐
智能助理
智能家居
五、未来发展趋势
更加智能化
更加人性化
更加普及化
六、总结
回顾 AI agent 的定义和特点
强调 AI agent 对未来的重要性

Auto-Prompt

LLM驱动的Prompt优化：我是一个演讲者，演讲的主题是AI agent；演讲为科普向，应简单明了，不应涉及过多技术细节，但要将囊括AI agent各方面的内容；演讲时长3分钟，大纲内容的多少应考虑语速与时长；演讲应具有故事和幽默的元素，使得听众更容易接受。请根据上述要求，帮我准备一个大纲。

↓

将上述prompt输入GPT并返回结果

Chain/Bot

在内部知识库中搜索有关问题：大纲的标准模板

↓

将知识库中“相关文字+原始输入”交给GPT

↓

GPT运行

↓

给出结果

Plugin插件

用户除输入提示词外，还增加，请搜索相关配图。

↓

LLM选择关键词

↓

调取图片搜索插件

↓

图片链接与回文输出

Agent 智能体

LLM驱动确认角色：我是一个演讲内容创作者；

↓

LLM驱动确认目标：准备个agent演讲大纲

↓

LLM驱动计划：
1. 了解agent
2. 准备大纲
3. 要考虑故事元素、时长节奏

↓

LLM驱动步骤1：
1.1 使用搜索引擎搜索“AI agnet”
1.2 对搜索结果进行汇总；
1.3 分析搜索结果内容

↓

LLM驱动步骤2：撰写大纲

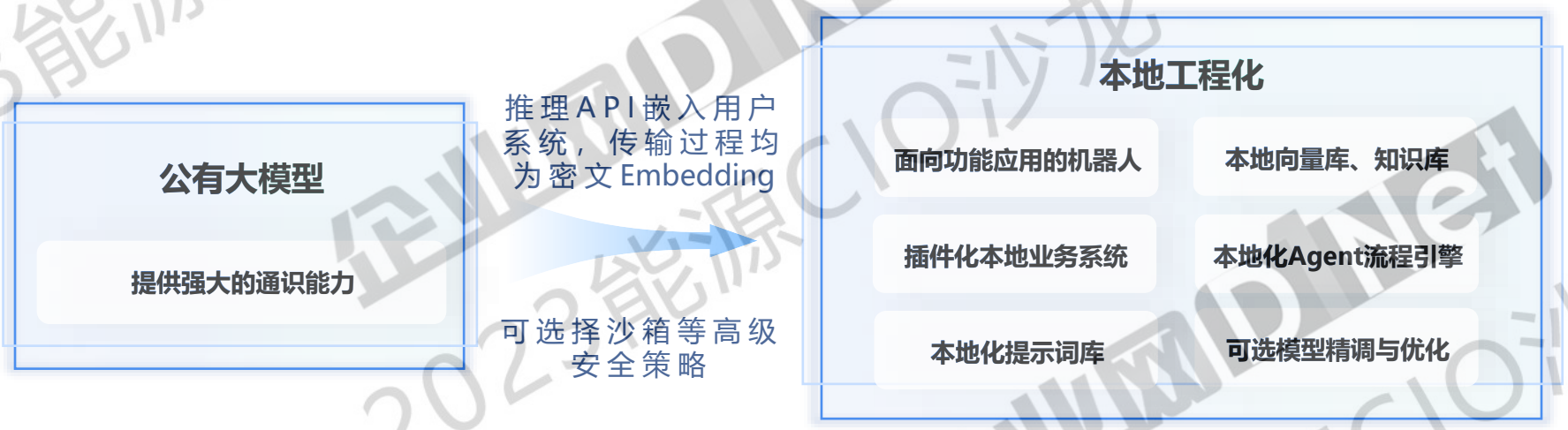
自己批改，重复了4次

一、引言
1. 介绍自己
2. 引起听众对 AI agent 的兴趣
二、什么是 AI agent
1. 简要解释AI agent的定义
2. 强调其在现实生活中的重要性
三、AI agent 的应用
1. 列举几个实际的例子，如智能家居、智能客服等
2.解释它们是如何工作的
四、AI agent 的发展
1. 强调 AI agent 技术正在不断发展
2. 展望未来可能的应用场景
五、结论
1. 总结 AI agent 的重要性
2. 鼓励听众对 AI agent 保持关注

数据隐私解决建议：本地工程化+大模型API

能源行业涉及国计民生，数据隐私与合规要求严格。

但自主训练大模型，成本高昂，GPU集群利用率低，模型更新慢。



基础设施与平台能力

面向大模型开发、服务的工具链、API和开放平台

基于大模型高性能工程化平台的资源调度

公有云、混合云

多样化、统一管理的PB级高性能数据存储层、高性能网络

火山引擎支持了国内诸多头部初创大模型

火山引擎

智谱华章

MINIMAX
新一代大语言模型

Kalends 昆仑万维

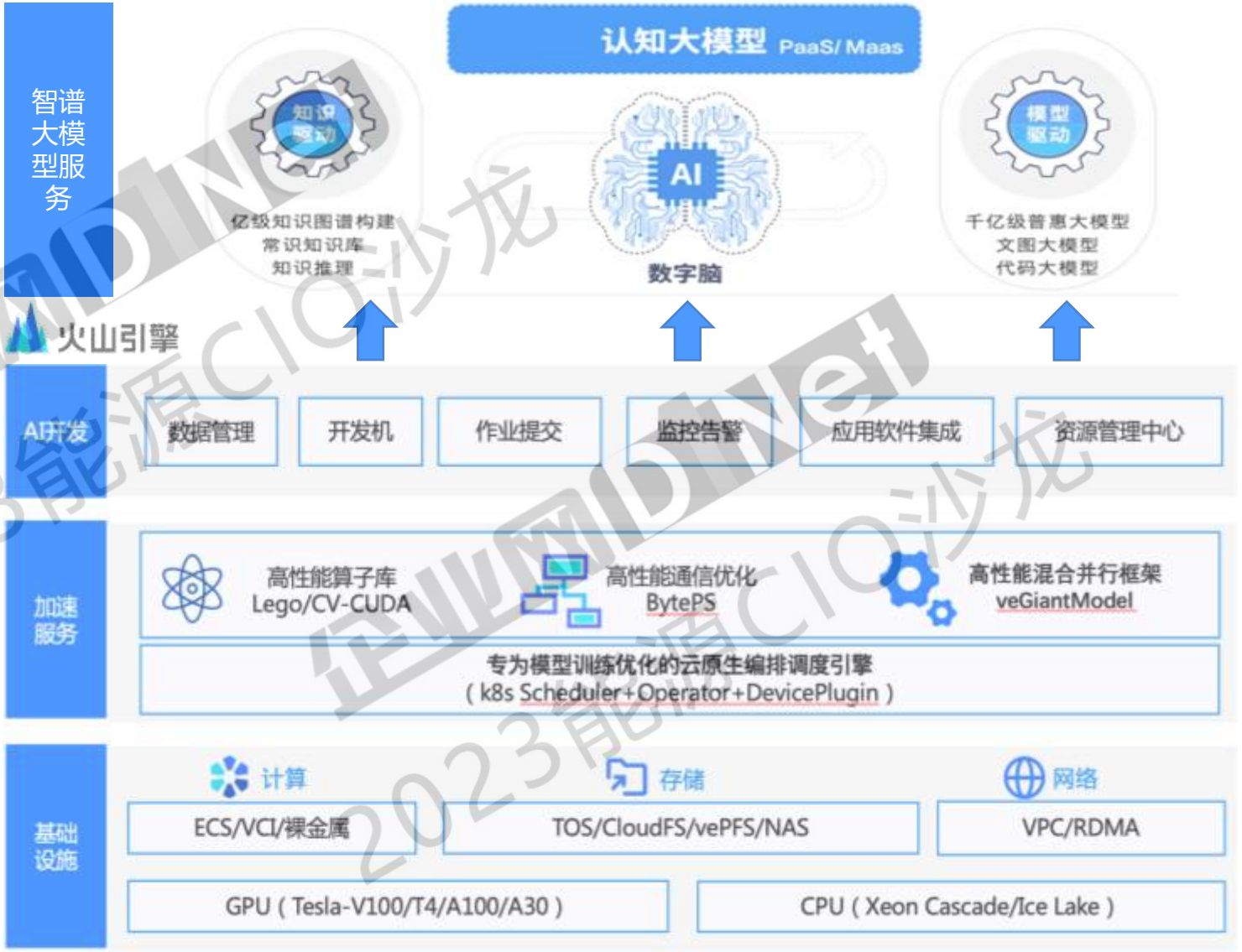
DriveGPT雪湖·海

痛点

1. 内存困境：100B的参数，40+张A100才能存放下。
2. 通讯挑战：常规网络环境带宽偏低的问题会引发通信瓶颈，限制了模型训练的效率
3. 性能瓶颈：大规模训练技术中，不仅要求AI芯片的计算性能足够强悍，同时也依赖AI框架大规模分布式训练的运行和调度效率。

解决方案

1. **超大算力池：**搭载英伟达 Tesla A100 80GB；2TB CPU Mem；单一集群 2000+ GPU 卡，提供 1 EFLOPS 算力。
2. **超强网络性能：**机内 600GBps 双向 NVLink 通道，800Gbps RDMA 网络高速互联，支持 GPU Direct Access
3. **并行文件系统 vePFS：**百 Gb 带宽，亚毫秒延迟，支持数亿小文件随机读取
4. **资源池化，排队调度：**降低资源闲置，加速模型训练



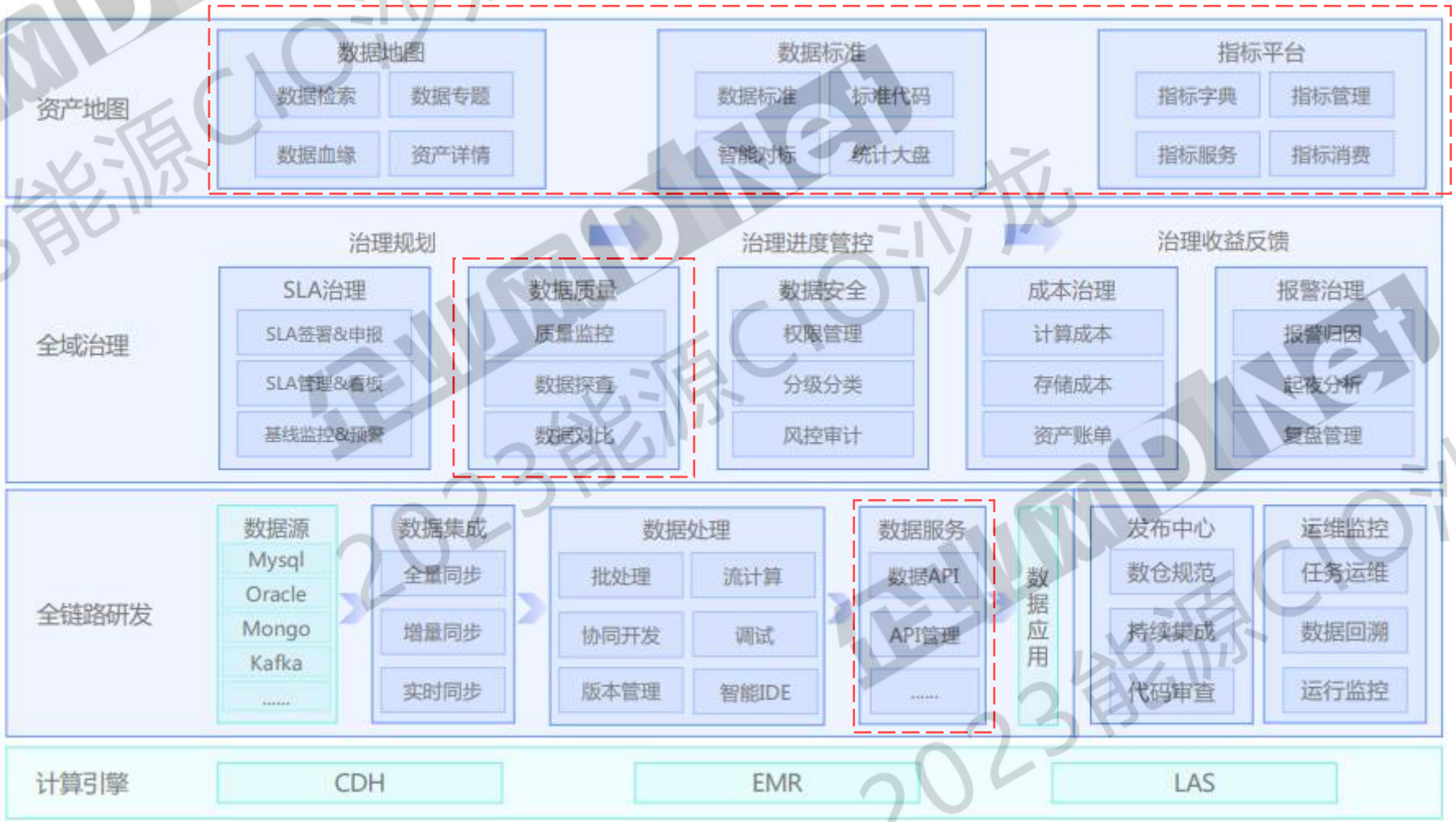
03 数字化的基石

面向智算的基础设施解决方案

火山引擎

“数据的生产与消费” —— 高效化数据治理管理与工具

火山引擎



“数据的生产与消费” —— 大模型应用

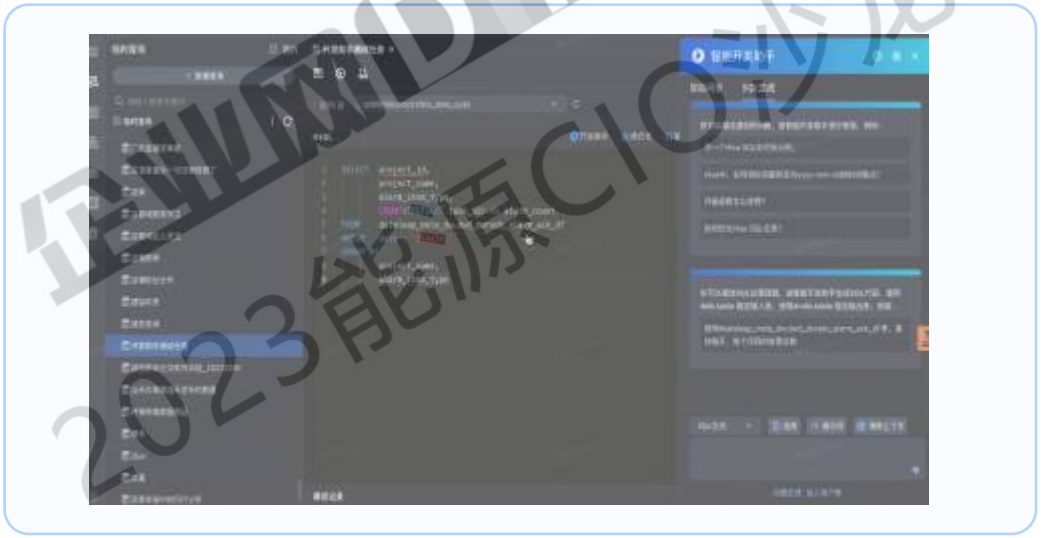
火山引擎

BI分析助手

智能数据洞察分析助手是一款基于大模型能力的自助分析智能助手，可以快速通过自然语言对话完成 SQL 查询、数据可视化查询与分析等一系列操作。围绕“提高生产力”而建设的分析助手，能以极低的使用门槛帮助更多业务人员快速自主进行 BI 分析，有效减少繁琐的取数与图表制作等工作，帮助人们缩短数据分析周期，提高生产力，让数据发挥价值 。

开发助手

实现通过自然语言描述，自动生成代码，针对已有的代码可以自动实现自动生成、修复，优化、解释与注释等。对话式方式进行文档搜索、函数使用、代码示例等问题咨询。助力平台用户减少基础开发工作量、提升开发效率。



找数助手

对话式的数据检索能力，解决用户找数据与用数据诉求。通过AI加持推动让搜索过程更聚焦。同时伴随模型语义理解能力的逐步提升，其全链路的检索效率更高，使得资产以低成本管理、促进自助式数据消费

MLP平台加速业务发展

云原生调度 —— 超大规模异构计算实践

火山引擎

节点数：120w+ 最大集群规模：2w+

在线微服务：30w+ 日变更数：3w 总存储量：10EB+

ByteDance 字节跳动

海量服务支撑	大规模调度	在线/离线混部
配置中心	单集群20000+节点	资源弹性调度，削峰填谷
全局网络与安全	镜像预热	K8s+yarn
Docker -> Containerd	集群联邦	离线任务：训练、视频转码
FaaS安全沙箱	弱网边缘计算	超售+服务弹性伸缩
对接DevOps	AML GPU调度	Quota计量

集群管理：K8S集群 | 服务容器化 | 异构计算 | 边缘计算 | 多活调度

资源域Quota：计算资源 | 存储资源 | 网络资源

云原生 PaaS 平台

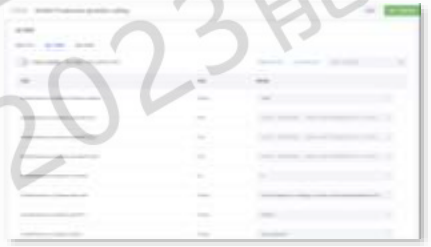
多云基础设施：Region 1 | Region 2 | Region 3 | Region 4

云原生调度 —— 贴近行业用户的产品化

火山引擎

SaaS化的硬件资源使用体验

可视化/模板



适合新手入门，只需要编辑参数即可投递计算任务

命令行/IDE

适合开发人员，通过命令行或IDE编辑脚本和投递作业

Application Aware Scheduling

性能策略
Performance

均衡策略
Balance

成本策略
Cost

基础设施篇—— 计算、网络、存储

火山引擎

RDMA

提供最高800GbpsRDMA高速互联网络，支持GPU Direct Access

vePFS存储

百GB带宽，亚毫秒延迟，数亿小文件轻松读取

超大规模

单一集群2000+GPU卡提供1EFlops算力

CPU

Intel、AMD、通用性、计算型、高主频型、大内存型

NVSWITCH

GPU机内600GBps双向NVSwitch通道

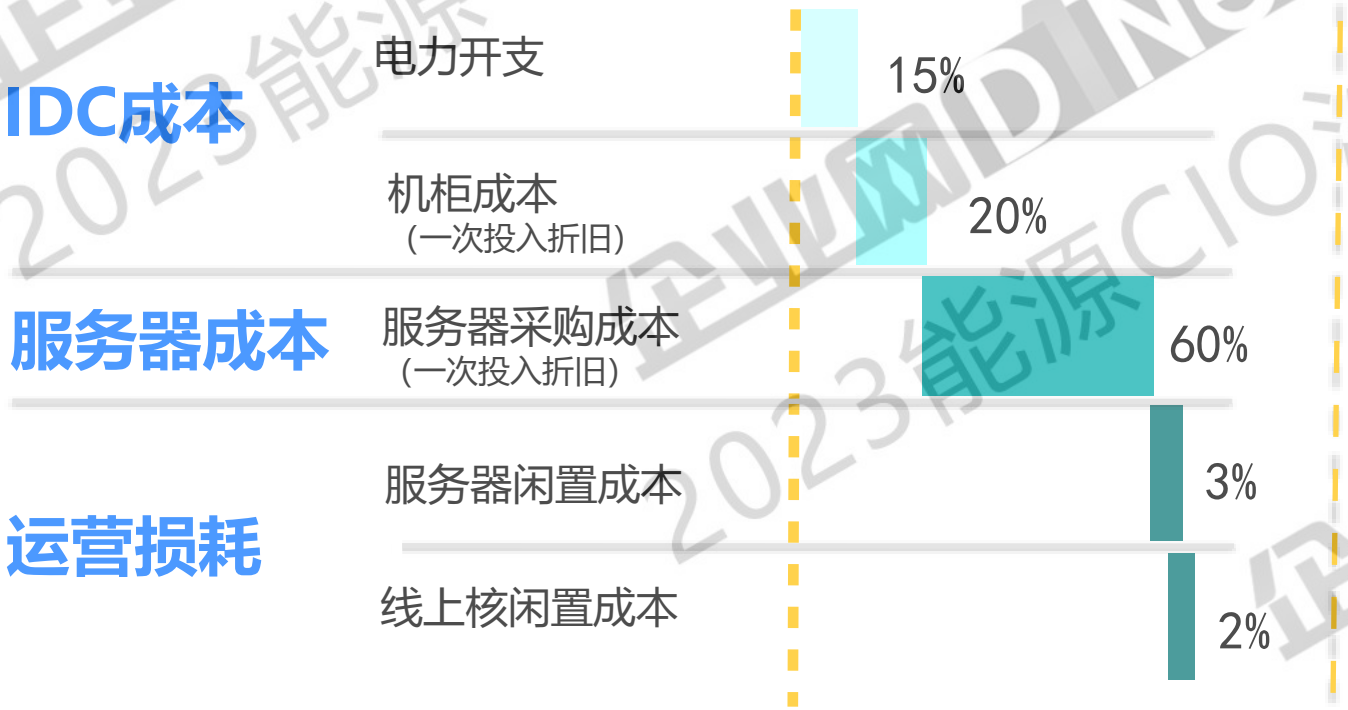
GPU

Nvidia训练与推理卡

算力基础设施中可变的是电力与运营

火山引擎

提供降低运营损耗的技术与支持



1 **规模**

- 依托字节集团百万级服务器体量的采购成本优势
- 硬件机型内外统一，提升火山引擎的供应效率

2 **技术**

- 持续优化的虚拟化底座与自研DPU，逐步将物理机损耗降低至0
- 大规模无感且无损的热迁移调度能力
- 基于字节海量业务场景的运维最佳实践，提升资源流转效率

3 **运营**

- 独有的内外并池能力，满足安全合规要求，极致压缩运营损耗

Thanks