AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·深圳 11.24-25

AI+
software
Development
Digital
summit

大规模云计算下节点故障预测AIops技术实践

马旭华 阿里云

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 深圳站

K+ 全球软件研发行业创新峰会

会议时间：2024.05.24-25

K+峰会 上海站

K+ 全球软件研发行业创新峰会

会议时间：2024.09.20-21

AiDD 峰会

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2023.11.24-25

AiDD峰会 北京站

AI+ 软件研发数字峰会

会议时间：2024.07.19-20

AiDD峰会 深圳站

AI+ 软件研发数字峰会

会议时间：2024.11.15-16

演讲嘉宾

马旭华

阿里云高级技术专家

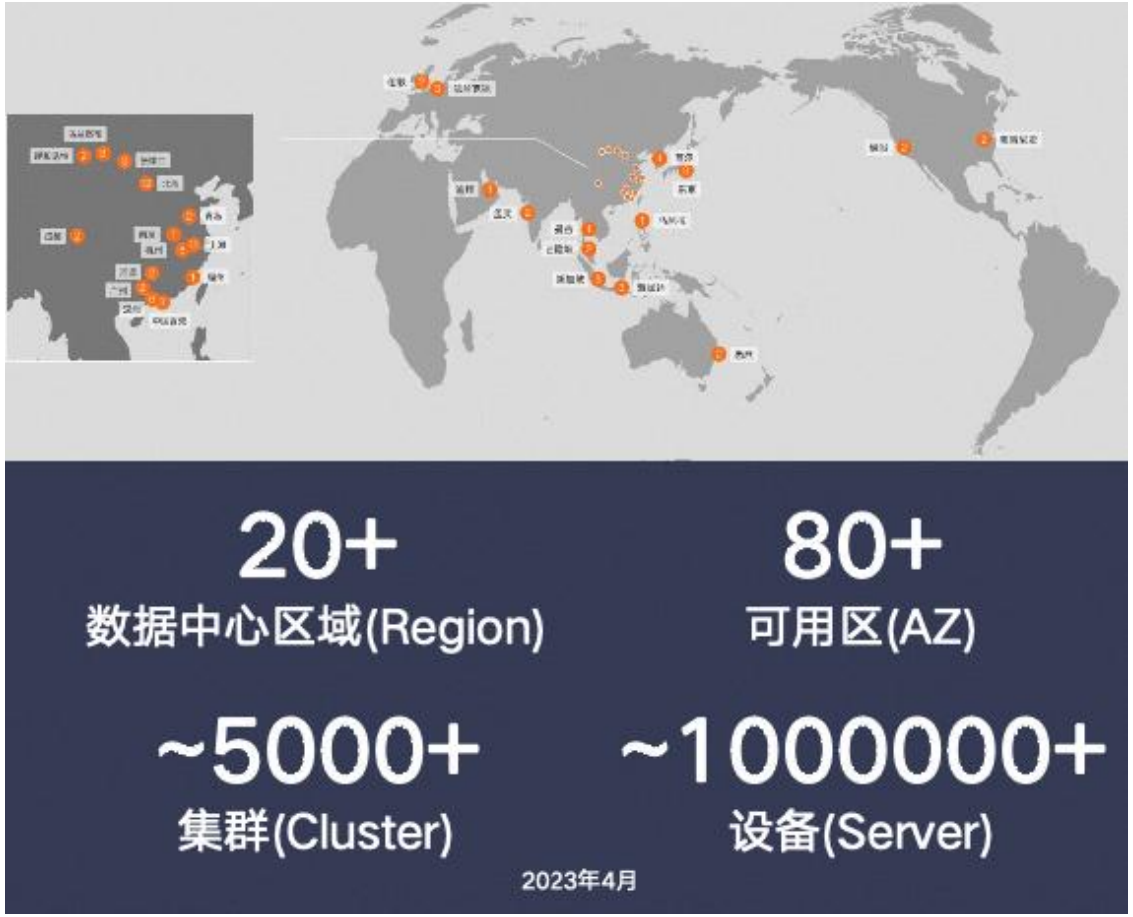
负责弹性计算产品的异常智能预测体系团队，负责AI算法在弹性计算产品稳定性体系的算法工程体系研发，聚焦于故障预测技术，受损感知，异常检测等领域的AIops系统研发

目录
CONTENTS

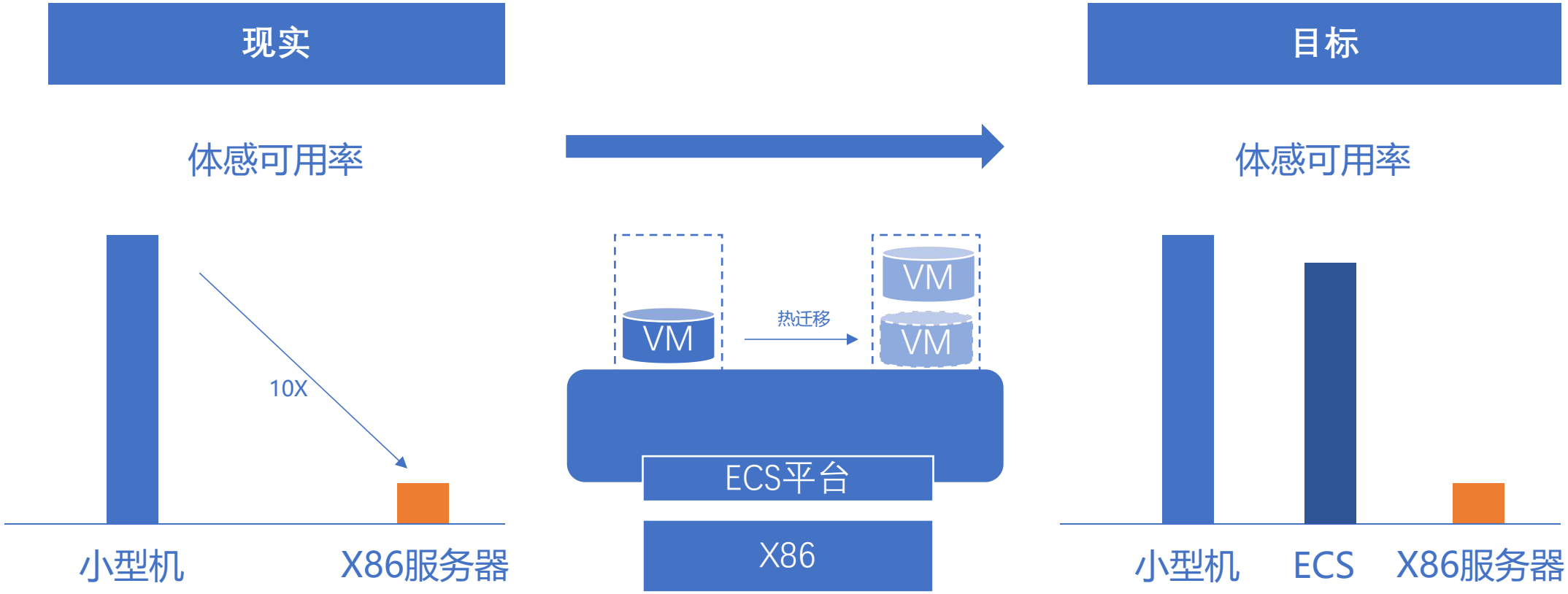
PART 01
大规模节点故障预测的背景&问题定义

背景 –弹性计算产品介绍

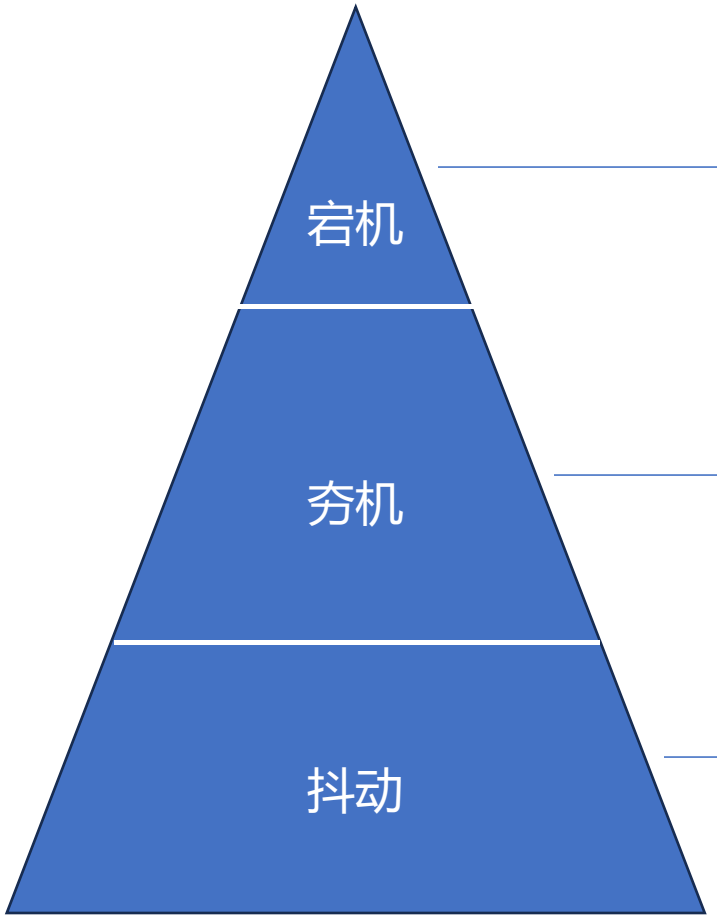
- 又名云服务器ECS(Elastic Compute Service)
- 云计算最核心基础IaaS服务之一
- 让大家像使用水、电、天然气等公共资源一样便捷、高效地使用服务器，实现计算资源的即开即用和弹性伸缩



ECS稳定性目标：用x86的硬件，提供小型机级别的稳定性



识别问题 –宕机，夯机，抖动



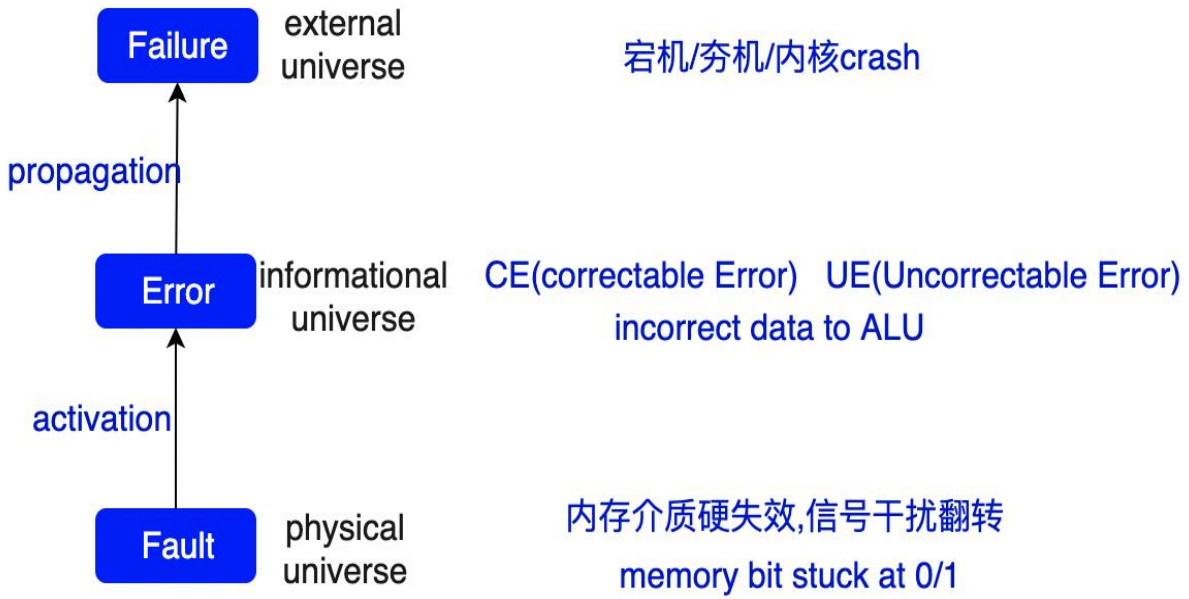
•现象:ECS资源100%不可用，多数因基础设施、服务器硬件或底层软件 原因导致。
•影响:所有未持久化的数据和配置都将丢失，该ECS实例上的业务将完 全中断。

•现象:ECS资源服务时断时续，甚至某些核心功能不可用或无法连接和 操作。如:OS 夯，IO hang等。
•影响:未持久化数据尚未丢失，但整个ECS几乎无法使用，有时甚至无法恢复、没有备份的机会。

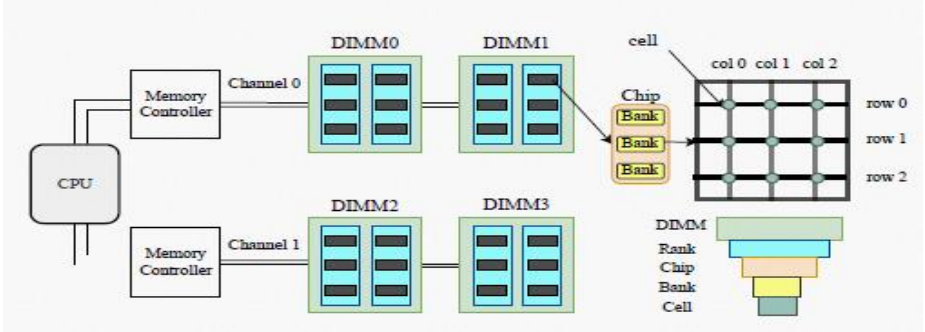
•现象:ECS资源核心服务可以正常使用，但在极端情况下会出现网络或 性能抖动。
•影响:着重影响抖动敏感用户，性能抖动可能导致用户压测等容量规划 付之东流，甚至可能因抖动引发用户应用系统雪崩效应，导致整体业务中断。

定义算法问题 –Fault/Error/Failure prediction

- Failure Prediction：节点Failure预测（宕机，夯机，性能受损）
- Error Prediction：UE（内存、PCIe、CPU） Prediction
- Fault prediction：硬件失效预测（内存，Disk Fault Prediction）



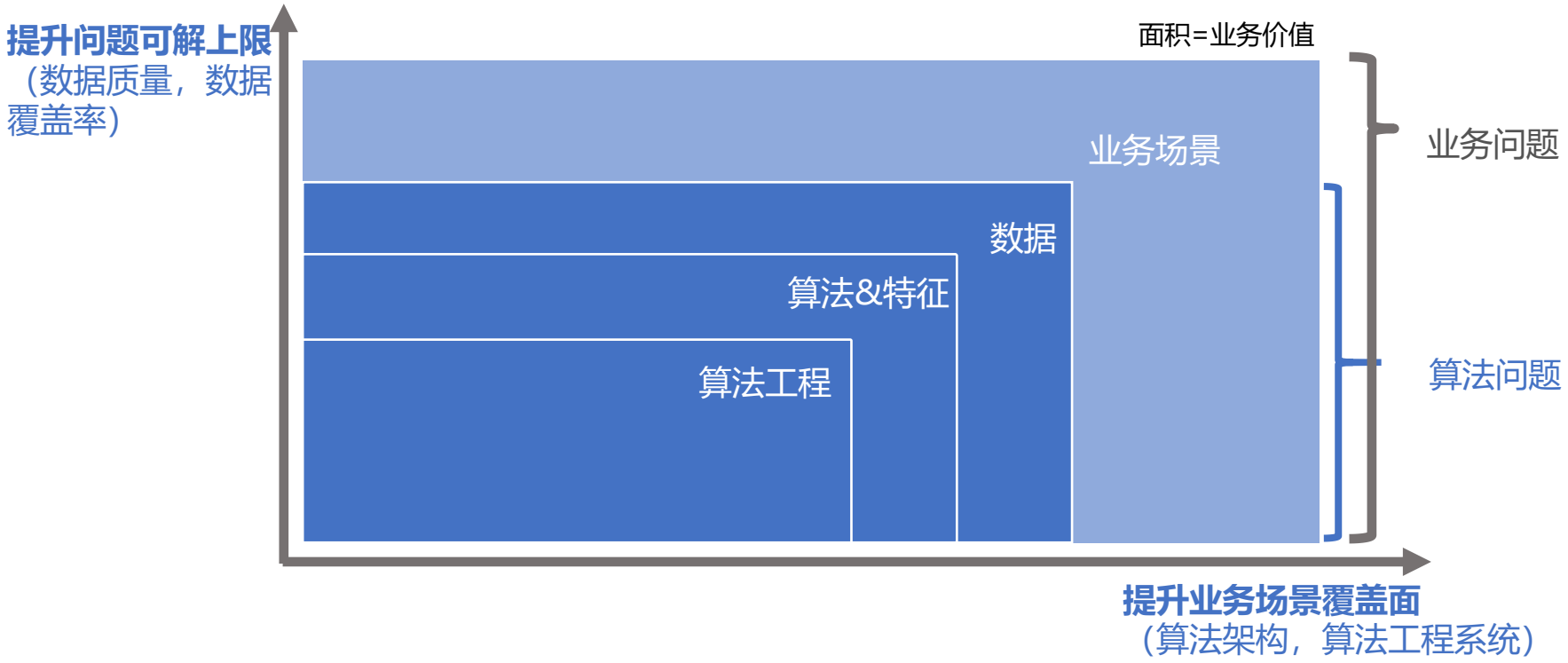
Fault prediction	Failure prediction
标签简单	标签难度大
静态	runtime
状态变化	突发性
实时性需求低	实时性要求高
硬件传感器数据	依赖full stack数据



PART 02

大规模节点故障预测问题（数据/算法/工程）

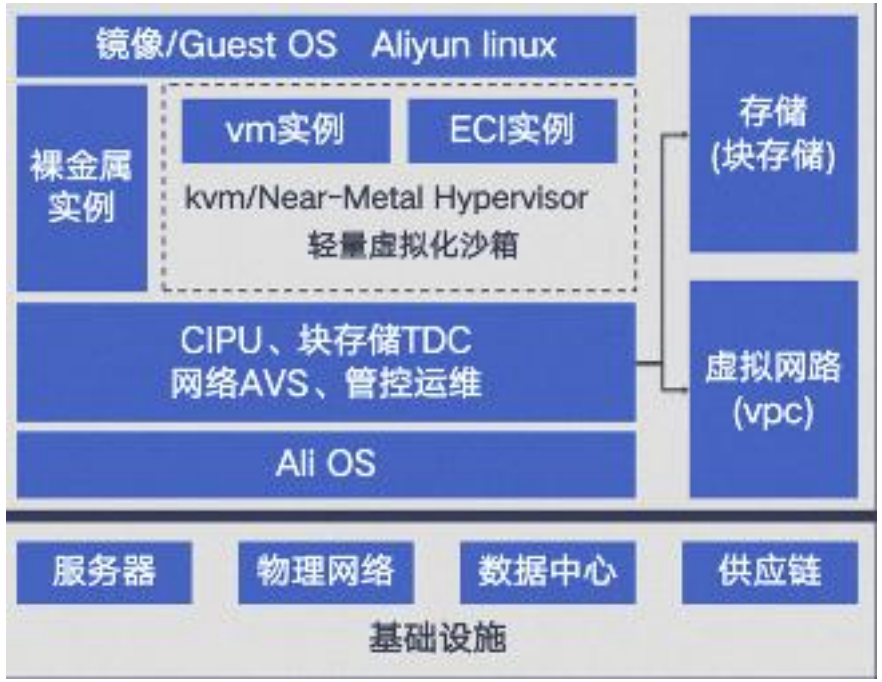
AIops工业落地需要解决的问题



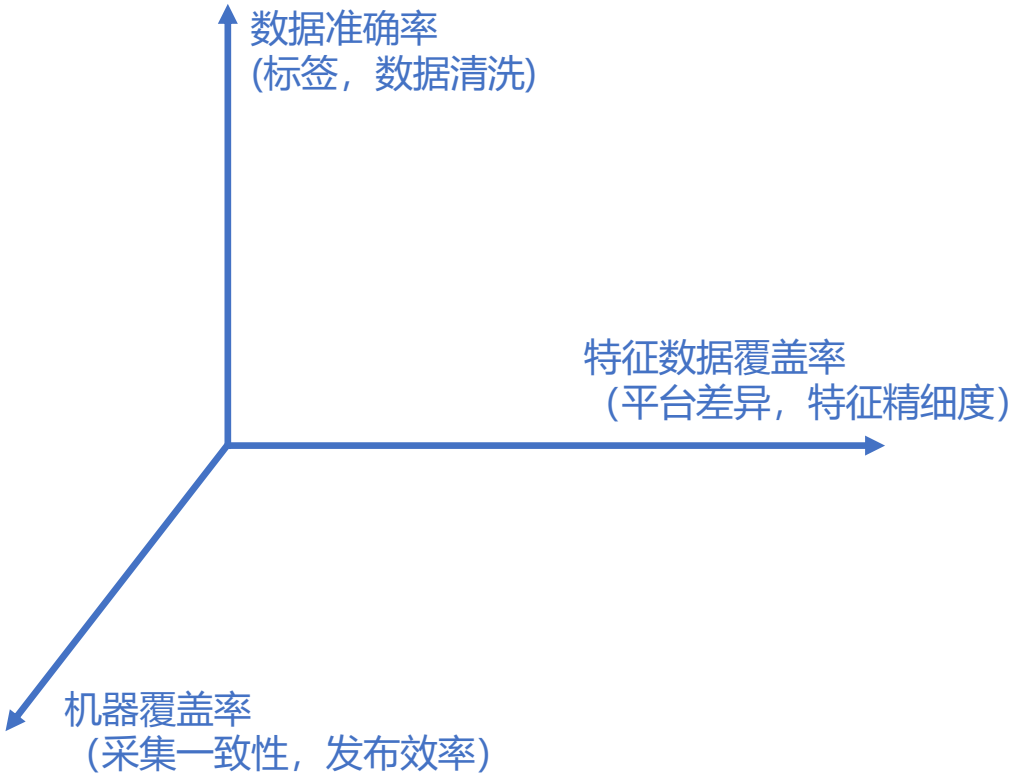
数据建设的问题与挑战：数据质量和复杂度

现状：业界无开源数据集，技术栈复杂，需要大规模环境下下长期积累

复杂度：横纵向技术栈



质量：算法“可用”的数据



算法面临的问题与挑战：适配故障预测问题的算法框架缺失

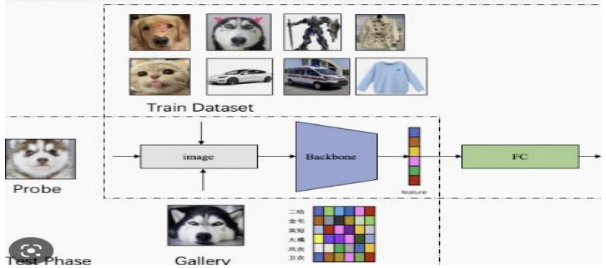
现状与问题：
- 改造问题与数据来适配算法框架
- 特征工程复杂度高/可复用性低
- 样本极度不均衡
- 算法可解释性

设计算法框架适配故障预测问题

NLP

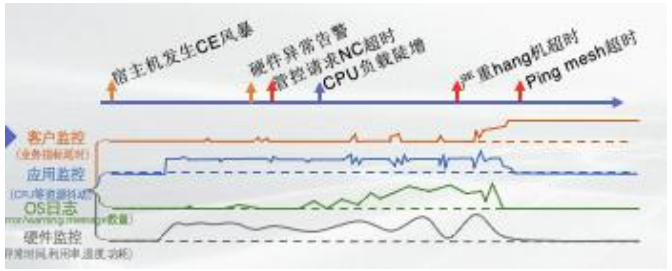
同质数据：单词
数据维度：一维序列
特性：局部相关性/远程相关性，位置敏感

图像处理



同质数据：像素
数据维度：2/3维
特性：局部相关性/平移/缩放不变性

节点故障预测



多模态泛时间序列预测问题
异质数据：单词/数值序列/异常特征
数据维度：多维(远>3维)
特性：局部相关性/远程相关性
多模态/时间敏感

算法工程系统的问题与挑战：实时性，数据污染，风险控制

- 大规模下预测实时性要求高 （延时需求，计算复杂性，规模）
- 算法迭代(模型退化)，上线的准确性评价（误预测结果污染标签导致性能衰退）
- “黑盒”模型，大规模运维风险控制

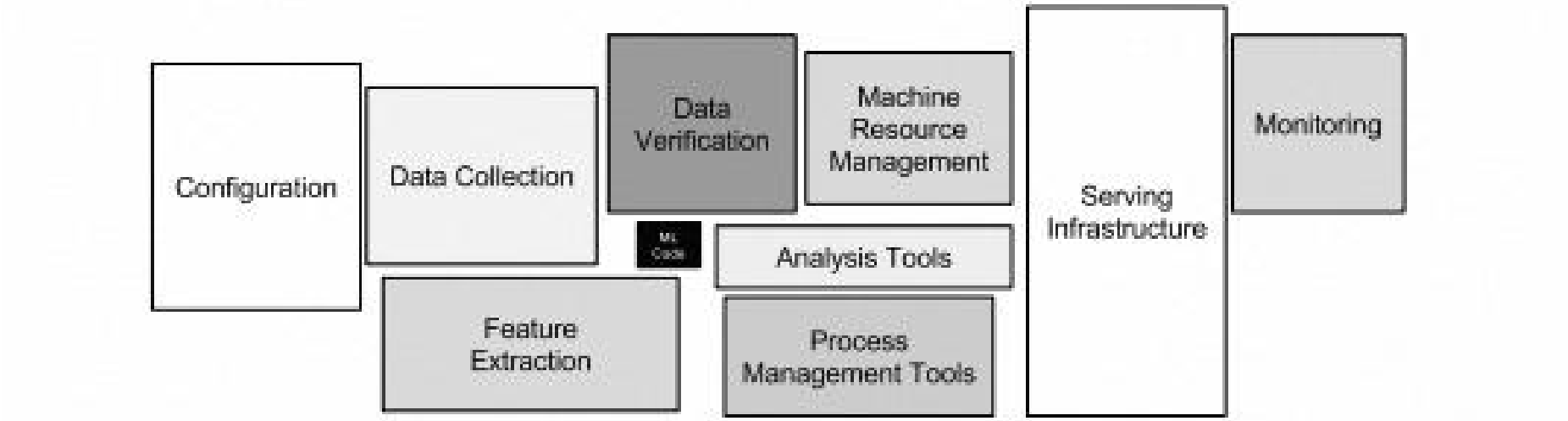
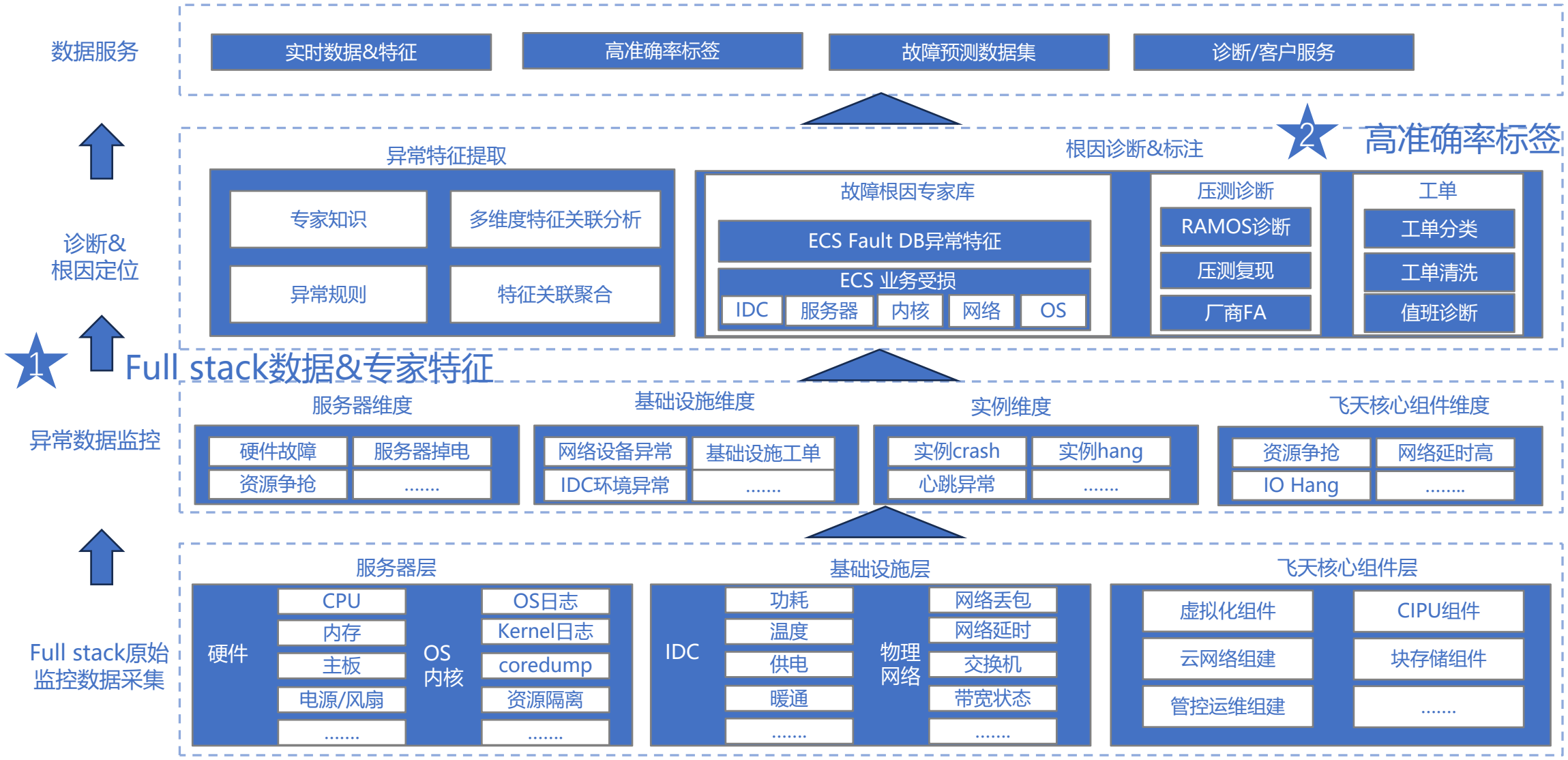


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Reference from “Hidden Technical Debt in Machine Learning Systems”

PART 03
大规模节点故障预测实践

数据采集系统--full stack数据&高准确率标签



数据采集系统—标准输出到软硬件协同数据定制

软硬协同数据上报技术体系—提升数据特征表达能力

- 软硬结合的数据定义&标准
- 更精细，更准，更快异常数据上报
- Full stack异常数据

VM

资源利用率 | 实例crash | 实例异常 | 实例抖动

虚拟化/内核

Error addr | 全栈性能监控数据 | 错误处理路径埋点(CE/UE) | 发布/变更

BIOS/BMC

流式数据上报 | Kernel first(CMCI,threshold ..etc)

硬件

CPU芯片定制(CE/MCA/ECC ..etc) | DDR4/DDR5 (on-die ECCetc)

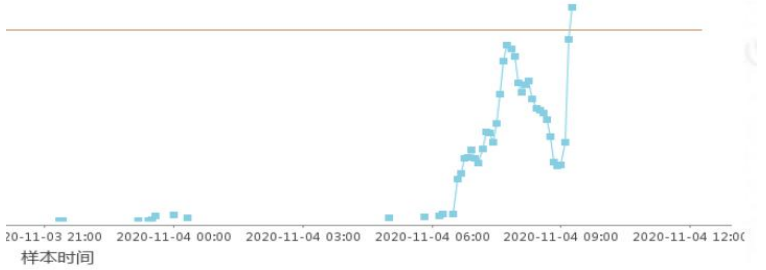
CIPU定制 | FPGA/主板 ...etc

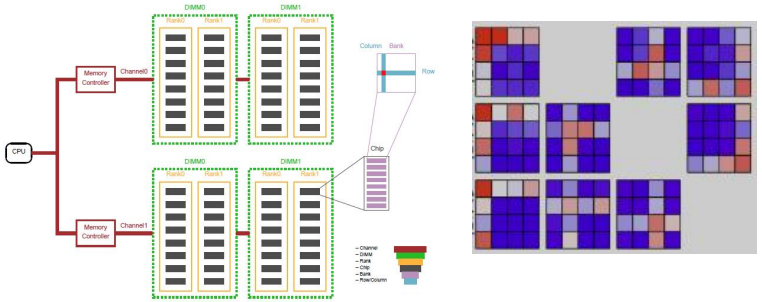
内存错误数据精细化示例

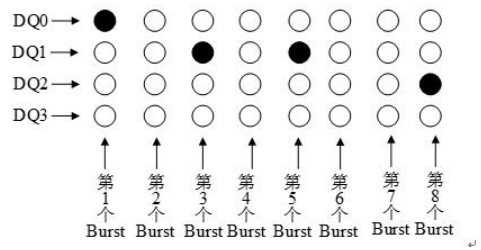
CE事件粒度

CE cell 粒度

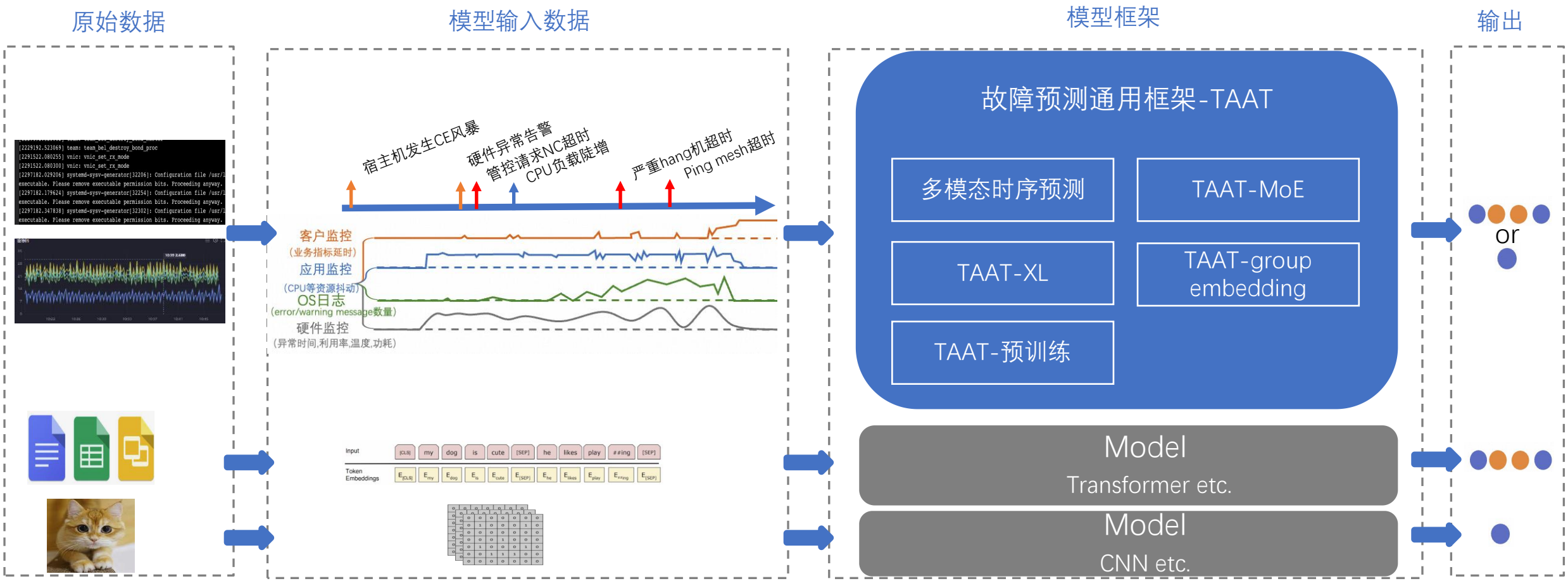
CE ECC bit 粒度



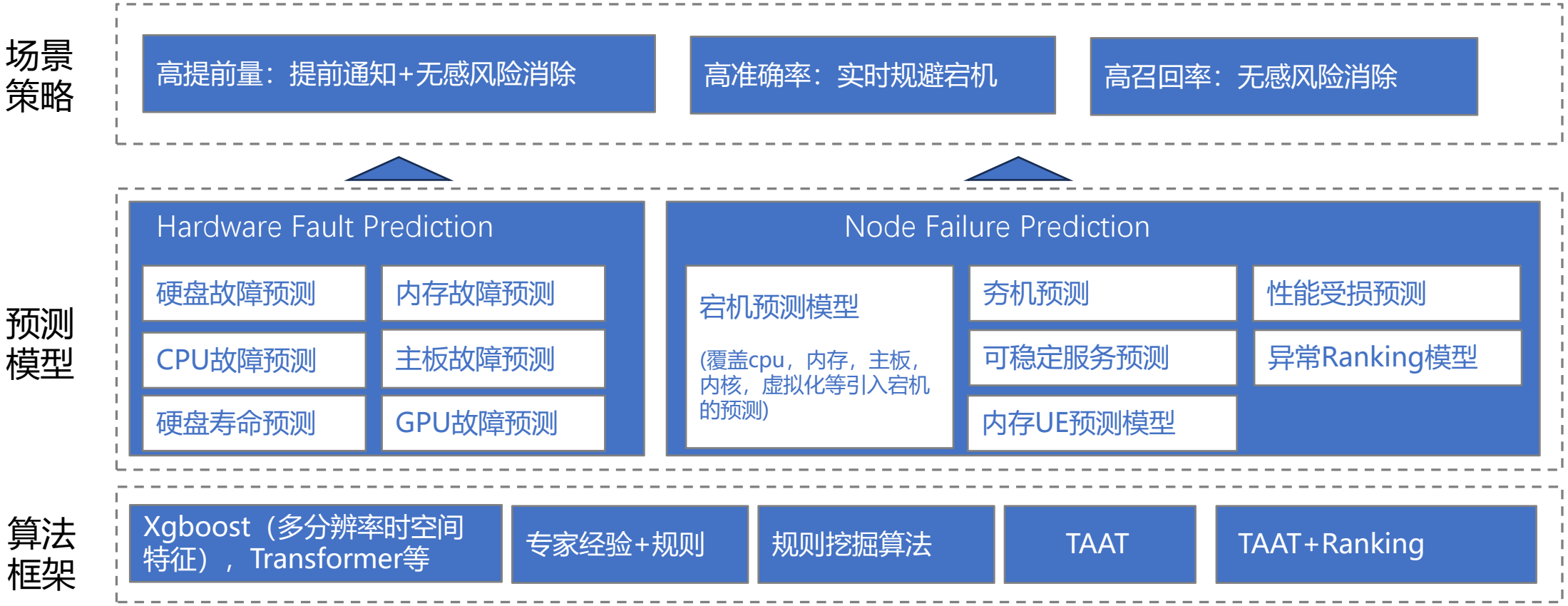




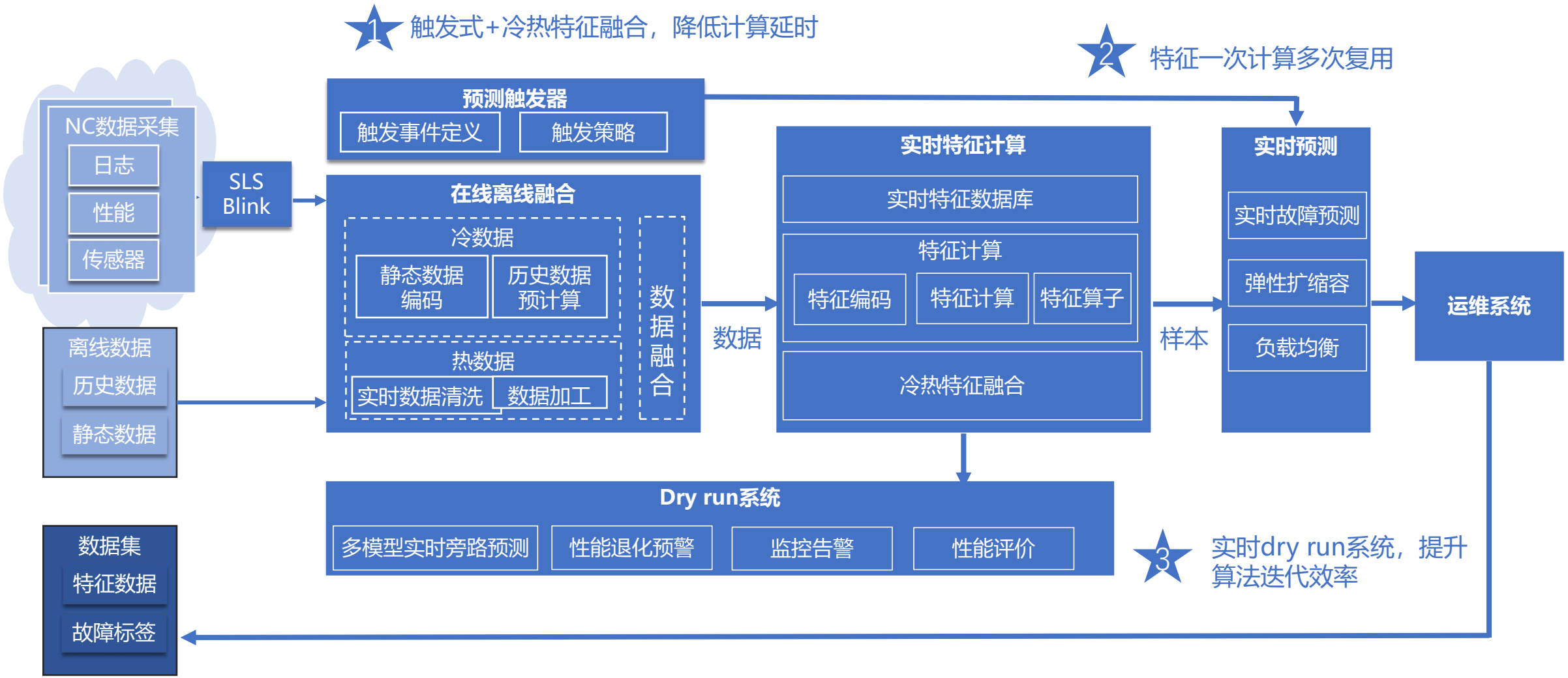
节点故障预测算法实践-自研算法架构



节点故障预测算法实践-模型框架



实时故障预测实践-实时数据-特征工程链路



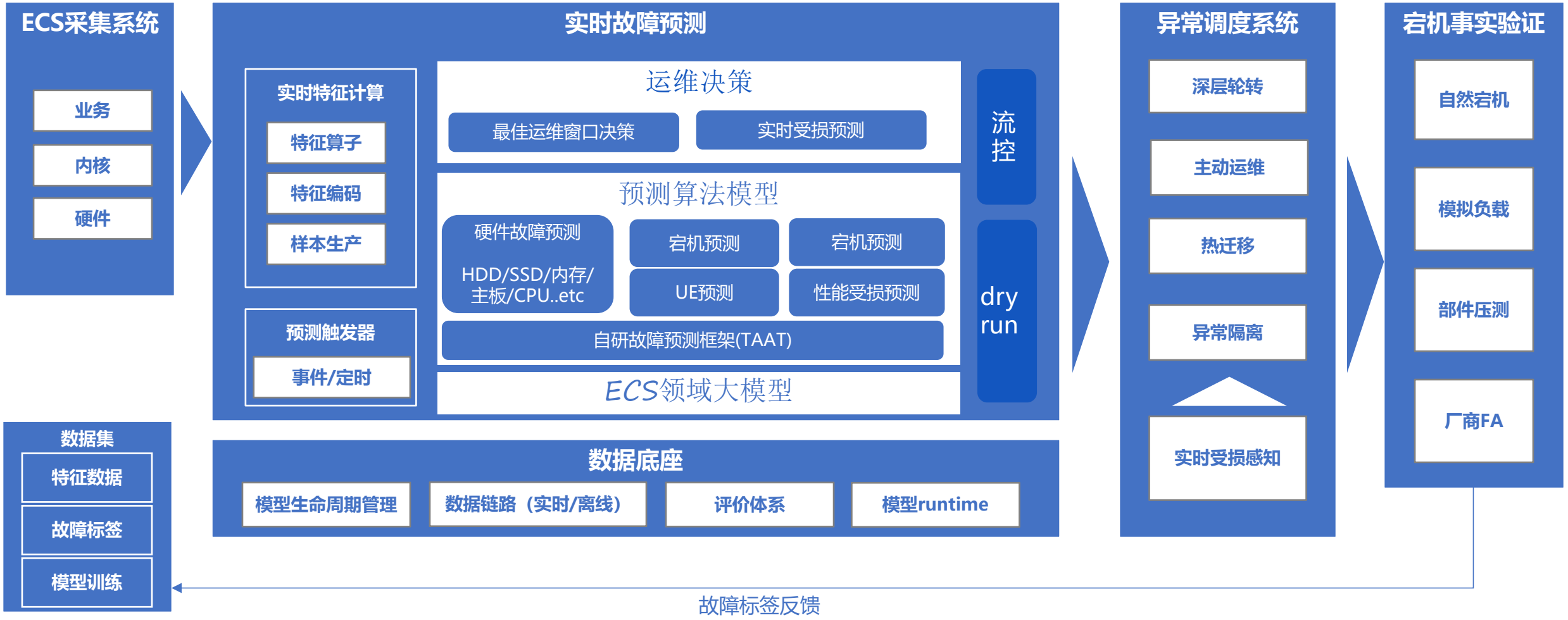
大规模节点故障预测系统实践—完整-自闭环故障预测体系

10年百万服务器精准打标

实时故障预测（完善的运维策略，完备的上线保护）

精确无感规避(实时受损预测与检测)

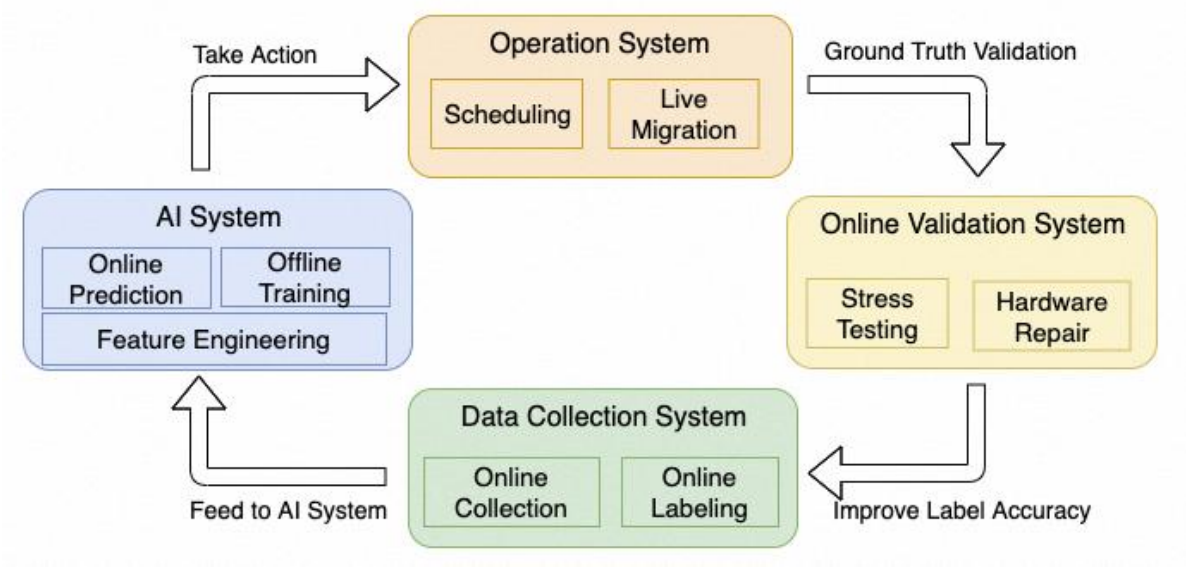
自闭环(持续迭代的基础)



PART 04
总结与展望

总结与展望

完整、自闭环的大规模云计算节点故障预测技术体系



展望

AIops算法体系优化

- 多模态泛时序预测算法-开放数据集
- 基于大模型的故障预测技术
- 实时故障预测算法效率的持续优化

软硬协同的故障预测技术

- 软硬协同的异常上报标准
- 基于软硬协同数据的故障预测技术

AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

THANKS