2025-05-21 AI峰会演讲

`powellli, 2025-05-19`

0. 背景和目标

AI峰会的AI Infra部分，长度10分钟，用于承接整体战略及强调Infra部分的升级能力。

1. 引入

尊敬的各位来宾，大家上午好！

我是腾讯云计算产品的李力，很高兴代表腾讯云AI Infra的产品体系在AI峰会上给大家分享，作为服务于腾讯自己以及外部大量AI用户和AI应用的底座，腾讯云智算体系的升级。

2. 新的挑战

前面Simon提到交互体验、模型能力和加速落地的三大趋势，非常清晰地阐明了当今AI发展的现状，对应对Infra的要求就是服务体验可靠性、训练速度和工具链三个层面上的能力要求。

今天我也会从高效、可靠和易用这三个方面给大家讲讲腾讯云智算的能力提升。

3. AI云原生与场景突破

我刚在台下看到很多我们的客户和媒体朋友们。大家应该还记得，腾讯云智算一直以来坚持的都是AI云原生的理念。

所谓AI云原生，就是以云原生的整套体系，支持AI Infra的发展。过去十几年，全世界的整个IT技术生态打造了以云为中心的基础设施最佳实践，这就是云原生的体系。我们在AI的浪潮中，坚持这一理念，在云原生的基础上提出了AI云原生的概念，也就是将云原生与AI的需求结合在一起，发挥云原生的标准化和技术生态的优势，将AI训练和推理应用更快更好地落地。

从AI云原生，场景上也从前两年的以LLM大模型训练为代表的场景延展到如今智能体、智能驾驶、机器人等更广泛更贴近实际应用的场景。

4. 高效、可靠和易用

正如前所述，AI计算的算力的需求、AI应用对上下游工具的需求，使得云智算平台在可靠、高效、易用上都有了更多的依赖。

5. 高效

模型的迭代速度 = 数据处理快 + 训练计算快
数据处理快：国内首个Serverless混合GPU调度平台，提交任务后自动调度智算资源，在数据处理场景下可实现10万并发和100万QPS
训练计算快：星脉自研通讯库相比业内领先的DeepEP还快30%，全新多机互联网络vRDMA接近无损拓展
自动驾驶案例：某头部车企智驾模型训练场景，通过云函数、数据万象进行

数据标注降本70%，提效50%。通过HCC+vRDMA进行训练任务，端到端性能提升30%。（比亚迪，名称已脱敏）

6. 可靠

保障沉浸式的AI应用交互体验，计算、存储、网络的稳定性缺一不可。以一个正常AI应用为例：
计算：星星海自研AI服务器针对AI场景定制优化，提高服务的稳定性。稳定性相比优化前提升了57%。
存储：基于自研的分布式读写和多集缓存机制，提高服务扩容的速度。以Deepseek满血版为例，加载速度从1小时缩短到20分钟。
网络：全球50+多节点接入，提高访问顺畅。网络丢包和抖动都降低了99%。
智能体案例：某手机头部厂商AI助手推理场景，通过GPU云服务器、GooseFS、CFS Turbo、多AZ等能力，稳定支持海量用户的丝滑访问。（荣耀，名称已脱敏）

7. 易用

智能体的加速落地离不开完备的工具链，比如知识库背后的向量检索技术、以及如何保障智能体的知识库数据不被入侵。

这些能力可直接被使用，也可通过智能体引擎PAAS被使用。

检索效率提升：向量数据库全新升级双路检索（向量+关键字）后召回率增加30%，同时支持千亿数据规模，500万QPS，性能业内TOP1。

服务异常感知：日志服务CLS原生支持AI应用、智能体的日志上报与分析，可实现3分钟快速发现异常。

安全合规保障：覆盖AI应用全生命周期的安全防护（数据处理->训练->推理->应用），实现小于2小时的安全响应处理时间，业内平均是20小时，相比业内减少90%。

传统搜索案例：某头部房产中介搜索场景，通过向量数据库双路检索提升了房屋租赁信息的召回率和检索效率，日志服务提供了日志托管提高30%定位效率。（贝壳，名称已脱敏）

8. AI原生云

收获国际权威机构和头部模型厂商认可。

Gartner亚太厂商第一，领先国内所有友商。

自研通讯库TRTM技术，被Deepseek在Github中高亮，并称其是“巨大进步”

9. 总结

全新的云智算能力，可同源同构延展到分布式云、专有云，让所有企业拥抱AI更简单。
也可同源同构延展到更多场景，比如自动驾驶、具身智能。腾讯云面向AI全场景，以做好深度洞察和充足准备。