伟仕佳杰
VSTECS
科技共建数智亚太

伟仕佳杰
打破AzureOpenAI 配额限制解决方案

伟仕佳杰集团

伟仕佳杰微讯

亚太地区领先的专注ICT行业解决方案的科技平台

www.vstecs.com

集团简介
GROUP PROFILE

伟仕佳杰控股有限公司（简称 “伟仕佳杰”）成立于1991年，2002年在香港主板上市，股票代码00856.HK。

伟仕佳杰是亚太地区领先的专注ICT行业解决方案的科技平台，是科技产品进入中国及东南亚市场的主要通路，为合作伙伴提供全品类、一站式、全链路的信息化服务，以提升运营效率，降低交易成本，推动科技企业数智化转型，加快信息化产业数智化进程。

企业实力
ENTERPRISE STRENGTH

- 1991年成立，专注深耕ICT行业
- 员工5,500名，认证工程师1,000+名
- 业务覆盖亚太9个国家
- 全亚太87个分支机构，50,000+渠道伙伴
- 香港主板上市（00856.HK），“深港通+沪港通”双标的

业务全景
BUSINESS LAYOUT

聚焦ICT行业，携手生态伙伴向千行百业提供产品和解决方案

通过**行业解决方案**深度服务千行百业
政府　金融　交通　制造　能源　教育　医疗　媒体...

通过**产业链解决方案**多元化链接ICT产业链上下游

工程师资源
产品　服务　技术解决方案

企业系统	消费电子	云 计 算	自研产品	全周期服务
基础建设 人工智能 计算 存储 软件 网络及安全 物联应用	智能终端 游戏娱乐 虚拟现实 整机外设及配件	分销转售 咨询迁移 运维管理 开发与增值	AI算力调度和运营平台 RightCloud多云管理平台 Vplus云平台	产品渠道赋能 服务合作赋能 解决方案赋能 技术支持赋能 系统集成赋能 培训认证赋能 市场营销赋能

AI产业链全覆盖　　云全生命周期管理　　信创产业深度布局

深度覆盖中国及东南亚9个国家

中国区500+认证工程师

基于在ICT领域的资源积累和长期实践，伟仕佳杰围绕人工智能、云计算、大数据、信息安全等领域，打造了一支技术领先、能力全面、行业经验丰富并拥有国际视野的专家团队，为上下游合作伙伴充分赋能。

AI算力与多云管理自研实力

佳杰云星有近百名聚焦软件研发和服务的技术工程师，拥有华为云HCIA/HCIP/HCIE认证、阿里云认证等云计算领域头部认证，及AI领域权威认证——人工智能工作级开发者认证HCCDP - AI。

AI基础设施领域技术能力

伟仕佳杰拥有华为、超聚变、昆仑、宝德、华鲲振宇和五舟等头部厂商售前、售后、投标工程师认证的工程师团队，为合作伙伴提供全周期的运维优化。聚焦高性能计算解决方案，团队具备超聚变、昆仑等FCS-Bidding、FCIE、FusionOne HCI FCS-Solution、KCIA、KCS-Pre-sales售前、服务器系统设计、解决方案实施等能力认证，能够根据用户需求提供定制化的技术支持和服务。

云计算领域技术能力

与VMware by Broadcom、华为云、AWS、阿里云、移动云、天翼云、百度智能云等国际顶尖厂商合作，为企业提供全面的云计算和大数据服务，在架构设计、专业咨询、服务集成、数据备份与灾难恢复等方面经验丰富。

C-]阿里云 AliCloud MVP 阿里云工程师最高认证	阿里云 AliCloud ACPACE 阿里云工程师认证	HCIA Huawei Cloud HCIA 华为云服务工程师认证	HCIE Huawei Cloud HCIP HCIE 华为云专家认证	vmware by Broadcom VMware VCP VMware 工程师认证	Microsoft Microsoft SQL DBA 微软数据库工程师认证
Microsoft Azure Azure Solutions Architect 微软解决方案架构专家认证	redhat. REDHAT RHCA 红帽系统架构师认证	ITIL Licensed Affiliate ITIL IT基础架构管理认证	PMI Project Management Institute PMP 项目管理资格认证	CCIE Cisco CCIE 思科互联网专家认证	PARTNER Advanced Tier Services • Distribution • AWS China Promotion • Small and Medium Business Services Competency • Migration and Modernization Services Competency

伟仕佳杰人工智能合作图谱

人工智能应用层

AI+泛安防
计算机视觉
大数据智能

AI+金融
营销客服
信贷风控
视觉产品

AI+政务
便民办公
政务大数据

AI+零售
运营优化
营销客服

AI+工业
安全生产
运维决策

AI+交通
智慧管理
自动驾驶

人机交互
对话式AI
消费级硬件

AIoT
产业级
消费级

AI+泛互联网
智能搜索问答
创作工具

AI+影视传媒
内容生成

AI+教育
教育工作
智慧校园

人工智能技术层

AI开放平台

AI开发平台

机器学习

知识图谱

自然语言处理

计算机视觉

智能语音

人工智能大模型层与工具层

通用基础大模型
大语言模型

视觉大模型

语音大模型

多模态大模型

垂直行业/领域大模型

大模型开放平台

AI Agents

AI开源社区

模型平台/模型服务

人工智能基础层

算力基础
智算软件平台
智能云服务
智算服务器
AI芯片

数据基础
AI基础数据网络
数据治理

打破OpenAI 配额限制

Azure OpenAI 配额限制

- Azure OpenAI 的配额限制策略主要基于订阅级别，按区域和模型类型分配 Tokens per Minute (TPM) 和 Requests per Minute (RPM) 限额
- 用户可将 TPM 灵活分配到多个部署中，总和不超过该区域的配额上限
- 不同的订阅类型和模型对应不同的配额

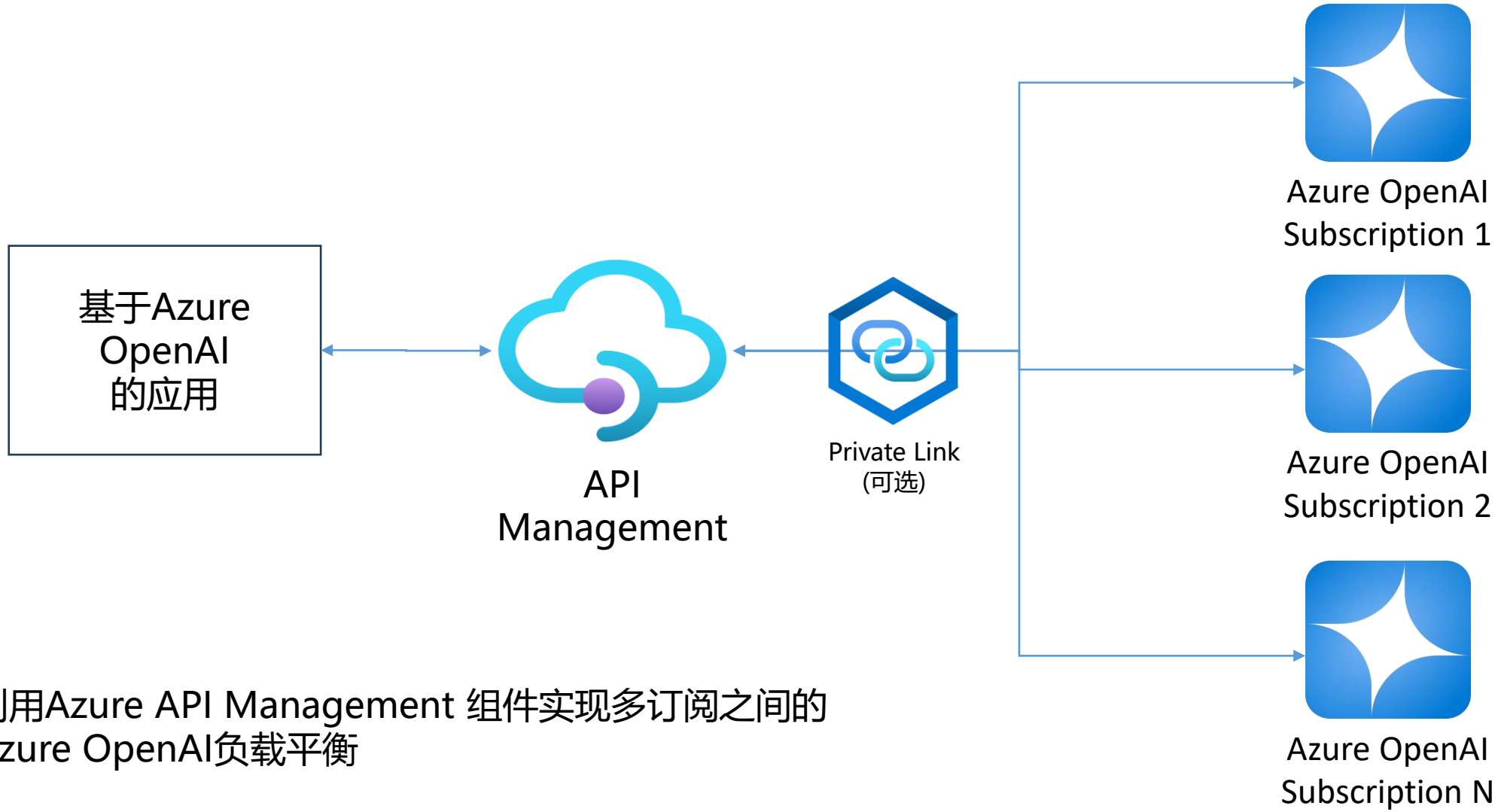
Model	Enterprise and MCA-E	Default	Monthly credit card-based subscriptions	MSDN subscriptions
gpt-4.1	5B	200M	50M	90K
gpt-4.1 mini	15B	1B	50M	90K
gpt-4.1-nano	15B	1B	50M	90K
gpt-4o	5B	200M	50M	90K
gpt-4o-mini	15B	1B	50M	90K
gpt-4-turbo	300M	80M	40M	90K
gpt-4	150M	30M	5M	100K
gpt-35-turbo	10B	1B	100M	2M
o3-mini	15B	1B	50M	90K
o4-mini	15B	1B	50M	90K
gpt-5	5B	200M	50M	90K

Microsoft Foundry 中的Azure OpenAI模型配额限制

由此引发的问题

- 面对大量并发时，调用端需要自己实现多订阅间的负载平衡
- 在单个订阅的TPM接近配额上限时，调用端会收到429错误
- 响应延迟增加和性能不稳定

解决方案



利用Azure API Management 组件实现多订阅之间的Azure OpenAI负载平衡

解决方案

- **扩展性和配额优化**：通过活跃-活跃负载均衡，将请求分布到多个订阅或部署（如预配吞吐量与标准部署），实现流量溢出和突发处理，突破单一订阅的 TPM/RPM 限额，支持多租户场景下的集中配额分配。
- **高可用性和可靠性**：支持故障转移、自动重试和断路器逻辑，当一个订阅因配额耗尽（429 错误）或故障时，智能路由到可用后端，确保服务连续性，并减少客户端代码负担。
- **性能提升**：采用优先级分组路由（如优先使用高优先级后端），避免传统轮询延迟，实现随机负载分布和无延迟切换，提高整体响应速度和吞吐量。
- **成本控制**：允许低配预配实例并用标准实例处理溢出，优化资源利用，避免过度配置导致的浪费，同时支持基于客户端的计费和展示模型。
- **安全与合规**：在网关层集中管理凭证终止、客户端识别路由和模型隔离，提供比 Azure OpenAI 实例级 IAM 更细粒度的访问控制，增强多订阅环境的安全性。
- **监控与可观测性**：统一收集跨订阅的遥测数据、日志和配额使用指标，便于仪表板可视化、警报设置和使用追踪，支持多区域冗余部署的单一控制平面管理。

负载平衡策略选择

- **Round-Robin（轮询）：静态算法，按顺序均匀分配请求到多个后端实例（如不同订阅的 OpenAI 部署），适用于基本负载均衡场景。**
- **Random（随机）：随机选择后端，避免单一实例过载，常用于简单流量分散。**
- **Priority-Based（基于优先级）：优先路由到高优先级后端（如高配额订阅），当其不可用（如 429 限流）时降级到次级，支持分组和权重内部分配，提升资源利用率。**
- **Weight-Based（基于权重）：自定义权重（如根据 TPM 配额比例分配流量），通过策略表达式实现动态路由，优化多订阅配额使用。**

API Management 配置选择

	Basic v2	Standard v2	Premium v2
适用场景	团队和项目的 API 管理	启动组织内部的 API 项目，并随着项目的发展逐步扩展。	适用于庞大请求量的企业级场景
基础价格	$150.01 per month8	$700/月	$2,801/月
扩展价格 (每增加一个单位)	$150.01/月	$500/月	$1,401/月
月请求量上限	基础价格内含10M次请求	基础价格内含50M次请求	无上限
	每增加1百万次请求$3	每增加1百万次请求$2.50	

利用生产环境数据确定APIM SKU

类别：Azure OpenAI - HTTP 请求

展开表

指标	REST API 中的名称	单位	集合体	尺寸	时间粒度	DS 导出
Azure OpenAI AvailabilityRate 使用以下公式计算可用性百分比：(调用总数 - 服务器错误数)/调用总数。 服务器错误包括任何 >=500 的 HTTP 响应。	`AzureOpenAIAvailabilityRate`	百分比	最小值、最大值、平均值	`ApiName`、`OperationName`、`Region`、`StreamType`、`ModelDeploymentName`、`ModelName`、`ModelVersion`	PT1M	否
Azure OpenAI 请求 一段时间内对 Azure OpenAI API 的调用次数。 适用于 PTU、PTU 管理的部署以及即用即付部署。 若要细分 API 请求，可以按以下维度添加筛选器或应用拆分：ModelDeploymentName、ModelName、ModelVersion、StatusCode（成功、客户端程序、服务器错误）、IsSpillover 以获取溢出信息、StreamType（流式处理请求和非流式处理请求）和作。	`AzureOpenAIRequests`	计数	总计（总和）	`ApiName`、`OperationName`、`Region`、`StreamType`、`ModelDeploymentName`、`ModelName`、`ModelVersion`、`StatusCode`、`IsSpillover`	PT1M	是的

[Azure OpenAI 监视数据参考]

我们能做什么？

- **架构设计与配额规划**：评估客户流量，设计最优的多订阅/多区域/PTU+PayGo 混合架构APIM
- **统一网关一键部署**：通过 ARM/Bicep/Terraform 快速交付支持优先级+权重+自动故障转移的完整策略模板
- **智能负载均衡策略实施**：Round-Robin、权重路由、429 自动降级、断路器、突发流量溢出处理
- **企业级安全与治理**：APIM 层集中密钥管理、客户端证书/JWT 验证、IP 白名单、订阅级模型隔离、Azure Policy 合规
- **统一监控与告警**：构建跨所有订阅的配额使用、延迟、429 错误率仪表板 + 自动告警与配额预警
- **自动扩缩容与成本优化**：按月/季分析使用率，动态调整订阅数量
- **企业支持与 SLA**：7×24 监控与应急响应，保证 99.95%+ 可用性，
- **迁移与落地陪跑**：从单订阅迁移到多订阅高可用架构，全程陪跑与知识转移

伟仕佳杰集团号

伟仕佳杰微讯号

伟仕佳杰

亚太地区领先的专注ICT行业解决方案的科技平台

www.vstecs.com