

云原生智能网关 购买指南



腾讯云

【 版权声明 】

©2013–2026 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

购买指南

云原生网关

价格说明

性能评估

AI 网关

价格说明

性能评估

版本生命周期管理

按量计费转包年包月

退费说明

欠费说明

购买指南

云原生网关

价格说明

最近更新时间：2026-07-23 17:07:31

本章节主要介绍云原生网关的产品版本和计费模式。云原生网关已于北京时间2022年8月28日正式开始商业化计费。

⚠ 注意：

- 自北京时间2022年8月28日起，云原生网关已结束免费公测，正式开始计费。
- 您的实例将会自动转为按量计费，如无需使用请及时销毁。如需长期使用，可转为包年包月模式。
- 除已商业化组件外，其他组件仍处于内测/公测免费阶段。

产品版本

网关按照版本分为专业版、标准版和开发版，三个版本的对比如下：

功能点	专业版	标准版	开发版
适用场景	适用于生产环境	适用于生产环境	适用于非生产环境、产品体验场景
节点规格	节点规格可选	节点规格可选	1核1G
节点数	节点数可选	节点数可选	单节点
实例类型	专享实例	专享实例	专享实例
高可用	多节点容灾	多节点容灾	不支持
自定义插件	支持	支持	支持
日志查询	支持	支持	支持
监控告警	支持	支持	支持
参数配置	支持	支持	支持
节点数变更	支持	支持	不支持
Kong Ingress	支持	支持	支持

Controller			
集群分组管理	支持	支持	不支持
节点自动扩缩容	支持	不支持	不支持
CLS 大盘	支持	不支持	不支持

计费模式

引擎实例费用

网关提供按量计费（后付费）和包年包月（预付费）两种购买方式。

- 按量计费模式：您可以随时开通/销毁实例，按实例的实际使用时长付费。计费时间粒度精确到小时，不需要提前支付费用，每天进行结算，可以有效避免资源浪费。
- 包年包月模式：即在新建实例时支付费用。适用于长期稳定的业务场景，费用较按量计费模式更为低廉。

引擎实例计费项包括引擎节点规格（CPU 和内存）和节点数。

按量计费（后付费）模式

在后付费模式下，请按需选择引擎规格和节点数，计费以小时为单位，不足1小时按照1小时计算。每小时进行一次结算，每24小时出一次账单，费用自动从腾讯云账号余额中扣除。

 **注意：**

以下引擎规格的收费标准针对的是一个节点，如果您的引擎配置多个节点，计费时需要乘以相应的倍数。

- 计费公式：**引擎实例费用 = 引擎规格单价 × 节点数 × 小时。
- 计费示例：**您选了 2核4GB 规格的标准版网关引擎，节点数为3，实例运行时间为2小时，那么总费用为 $0.8 \times 3 \times 2 = 4.8$ 元。

在您购买按量计费实例后，系统会为您分配相关的实例资源，无论您是否实际使用该实例及相关资源，都会按小时结算，产生计费。如果无需再使用可以操作释放，避免持续产生费用。

标准版		
产品版本	规格	单价（元/节点/小时）
标准版	1核2GB	0.7
标准版	2核4GB	0.8
标准版	4核4GB	1.3
标准版	4核8GB	1.6

标准版	8核16GB	3.2
标准版	16核32GB	6.4

开发版

产品版本	规格	单价（元/节点/小时）
开发版	1核1GB	0.65

专业版

产品版本	规格	单价（元/节点/小时）
专业版	1核2GB	0.9
专业版	2核4GB	1.1
专业版	4核4GB	1.67
专业版	4核8GB	2.1
专业版	8核16GB	4.2
专业版	16核32GB	8.3

包年包月（预付费）模式

在包年包月（预付费）模式下，请按需选择引擎规格和节点数，计费以月为单位，费用自动从您的腾讯云账号余额中扣除。

注意

以下引擎规格的收费标准针对的是一个节点，如果您的引擎配置多个节点，计费时需要乘以相应的倍数。

- **计费公式：**引擎实例费用 = 引擎规格单价 × 节点数 × 月。
- **计费示例：**您选了 2核4GB 规格的标准版 网关 引擎，节点数为3，实例运行时间为1个月，那么总费用为 $418 \times 3 \times 1 = 1254$ 元。

标准版

产品版本	规格	单价（元/节点/月）
标准版	1核2GB	367
标准版	2核4GB	418
标准版	4核4GB	671
标准版	4核8GB	836
标准版	8核16GB	1672
标准版	16核32GB	3344

开发版

产品版本	规格	单价（元/节点/月）
开发版	1核1GB	300

专业版

产品版本	规格	单价（元/节点/月）
专业版	1核2GB	477.1
专业版	2核4GB	543
专业版	4核4GB	872
专业版	4核8GB	1087
专业版	8核16GB	2173
专业版	16核32GB	4347.2

公网流量费用

公网流量费用支持按使用流量计费和按小时带宽计费两种模式。

按使用流量

按使用流量主要为下行流量，即按照网关返回客户端时产生的公网出流量进行计费。付费模式为后付费，每小时结算一次。

计费价格

地域	价格（单位：元/GB）
中国大陆地区（不含港澳台地区）、新加坡、雅加达、首尔、东京、法兰克福	0.80
中国香港、圣保罗	1.00
曼谷、硅谷、弗吉尼亚、孟买	0.50

按小时带宽

计费价格

带宽	单价（元/Mbps/小时）
≤5Mbps	0.044
>5Mbps	0.16

计费示例： 假设您选择了按小时带宽计费模式,购买了6Mbps的固定带宽，则总费用为： 5Mbps × 0.044 元/Mbps/小时 + （6Mbps – 5Mbps） × 0.16元/Mbps/小时 = 0.38元/小时。

支持的地域

下表为云原生网关所支持的地域信息：

地域	取值
中国大陆（广州）	ap-guangzhou
中国大陆（北京）	ap-beijing
中国大陆（上海）	ap-shanghai
中国大陆（成都）	ap-chengdu
中国大陆（南京）	ap-nanjing

中国大陆（上海金融）	ap-shanghai-fsi
中国大陆（深圳金融）	ap-shenzhen-fsi
亚太地区（中国香港）	ap-hongkong
亚太地区（新加坡）	ap-singapore
亚太地区（曼谷）	ap-bangkok
亚太地区（雅加达）	ap-jakarta
亚太地区（首尔）	ap-seoul
亚太地区（东京）	ap-tokyo
欧美地区（弗吉尼亚）	na-ashburn
欧美地区（硅谷）	na-siliconvalley
欧洲地区（法兰克福）	eu-frankfurt

性能评估

最近更新时间：2026-07-23 17:07:31

云原生网关支持按节点规格和数量购买，也支持在购买后调整节点规格和数量。本文介绍云原生网关 Kong 的性能指标，请根据业务需求选择合适的资源大小。

网关运行水位

网关运行水位说明如下：

- 推荐水位：CPU 使用率为30%，内存使用率为30%。
- 警戒水位：CPU 使用率为60%，内存使用率为60%。

在日常运行中，建议将资源使用率控制在30-60%，既可以保持较高的资源利用率，又可以应对请求突增的情况。在警戒水位下，网关可以获得完整的 SLA 保障。当运行水位超过警戒水位时，网关进入高负载状态，请求成功率和耗时可能受到影响，建议配置告警。

网关容量指标

节点规格和数量决定网关容量，增大节点规格或者增加节点数量可以提高网关容量。如果节点规格不变，随着节点数量的增长，网关容量基本保持线性增长。网关容量指标说明如下：

- 连接数：网关支持的最大连接数量
- HTTP 每秒新建连接数：网关支持每秒新建 HTTP 短连接的最大值。
- HTTPS 每秒新建连接数：网关支持每秒新建 HTTPS 短连接的最大值。

节点规格		2核4G	4核8G	8核16G	16核32G
节点数量		1节点	1节点	1节点	1节点
CPU 使用率	推荐水位	30%	30%	30%	30%
	警戒水位	60%	60%	60%	60%
内存使用率	推荐水位	30%	30%	30%	30%
	警戒水位	60%	60%	60%	60%
连接数	推荐水位	24000	48000	96000	192000
	警戒水位	48000	96000	192000	384000
HTTP 每秒新建连接数	推荐水位	2000	4000	8000	16000
	警戒水位	4000	8000	16000	32000

HTTPS 每秒新建连接数	推荐水位	600	1200	2400	4800
	警戒水位	1200	2400	4800	9600

网关性能数据

测试场景：客户端发送请求给网关，网关转发到后端服务器，后端服务器直接返回响应。

测试条件：不同的请求类型、连接类型和响应大小。

测试结果：客户端请求 QPS 如下所示，请求平均耗时小于 5 毫秒。

CPU 和内存使用率处于推荐水位（30%）										
协议类型	连接类型	响应大小	2核4G		4核8G		8核16G		16核32G	
			3节点	5节点	3节点	5节点	3节点	5节点	3节点	5节点
HTTP	短连接	512B	6500	10800	13000	21600	25000	41650	50000	83300
		1KB	5800	9600	11600	19200	23200	38650	46400	77300
		10KB	4500	7500	9000	15000	20000	33300	40000	66650
	长连接	512B	7000	11600	14000	23200	28800	48000	57600	96000
		1KB	6700	11100	13400	22200	27000	45000	54000	90000
		10KB	6000	10000	12000	20000	24000	40000	48000	80000
HTTPS	短连接	512B	1800	3000	3600	6000	7800	13000	15600	26000

		1KB	1700	2800	3400	5600	7500	12500	15000	25000
		10KB	1600	2600	3200	5200	7100	11800	14200	23650
	长连接	512B	6450	10750	12900	21500	26000	43300	52000	86650
		1KB	6200	10300	12400	20650	25200	42000	50400	84000
		10KB	5500	9150	11000	18300	22000	36650	44000	73300

AI 网关

价格说明

最近更新时间：2026-07-23 17:07:31

AI 网关产品购买链接，可点击后进入购买页。计费由**实例费用**、**数据处理费用**和**公网流量费用**组成。（产品化支持水平、垂直无损扩容）
标准版实例支持所有产品能力，仅在不同规格的实例版本中对可使用资源数量做限制。

计费方式

- **按量付费（后付费）**：计费以小时为单位，不足1小时按照1小时计算。每小时进行一次结算，每24小时出一次账单，费用自动从账户余额中扣除。具体出账时间以实际为准。
- **包年包月（预付费）**：计费以月为单位，包年按照12个月计算。费用自动从账户余额中扣除。具体出账时间以实际为准。

计费项

实例费用

- 按实例规格阶梯定价，实例规格分为：**节点规格**和**节点数** 两个维度。
- 节点规格：1C2G、2C4G、4C4G、4C8G、8C16G、16C32G。
 - 节点数：2 节点起售。

引擎实例计费项包括引擎节点规格（CPU 和内存）和节点数。

功能版本	按量计费（元/小时）	包年包月（元/月）
标准资源版本 1C2G	4.45	1,600
标准资源版本 2C4G	8.35	3,000
标准资源版本 4C4G	以控制台为准	以控制台为准
标准资源版本 4C8G	16.4	5,900
标准资源版本 8C16G	31.65	11,400
标准资源版本 16C32G	62.2	22,400

数据处理费用

在一个计费周期内，数据处理量为请求和响应的数据量之和。
国内站单价：**0.036 元/GB/小时**

公网流量费用

公网流量费用支持按使用流量计费和按小时带宽计费两种模式。

按使用流量

按使用流量主要为下行流量，即按照网关返回客户端时产生的公网出流量进行计费。付费模式为后付费，每小时结算一次。

地域	单价（元/GB）
中国大陆地区（不含港澳台地区）、新加坡、雅加达、首尔、东京、法兰克福	0.80
中国香港、圣保罗	1.00
曼谷、硅谷、弗吉尼亚、孟买	0.50

按小时带宽

带宽	单价（元/Mbps/小时）
≤5Mbps	0.044
>5Mbps	0.16

计费示例：假设您选择了按小时带宽计费模式，购买了 6Mbps 的固定带宽，则总费用为：

$5\text{Mbps} \times 0.044\text{元/Mbps/小时} + (6\text{Mbps} - 5\text{Mbps}) \times 0.16\text{元/Mbps/小时} = 0.38\text{元/小时}$ 。

性能评估

实例规格（2节点）	QPS（CPU 30%水位）	客户端连接数（CPU 30%水位）	QPS（CPU 60%水位）	客户端连接数（CPU 60%水位）
1C2G	1,500	20,000	3,000	40,000
2C4G	3,000	40,000	6,000	80,000
4C4G	以控制台为准	以控制台为准	以控制台为准	以控制台为准
4C8G	6,000	80,000	12,000	160,000
8C16G	12,000	160,000	24,000	320,000
16C32G	24,000	320,000	48,000	640,000

性能评估

最近更新时间：2026-07-23 17:07:33

AI 网关支持按节点规格和数量购买。本文介绍 AI 网关的性能指标，请根据业务需求选择合适的资源大小。

网关运行水位

网关运行水位说明如下：

- 推荐水位：CPU 使用率为30%，内存使用率为30%。
- 警戒水位：CPU 使用率为60%，内存使用率为60%。

在日常运行中，建议将资源使用率控制在30-60%，既可以保持较高的资源利用率，又可以应对请求突增的情况。在警戒水位下，网关可以获得完整的 SLA 保障。当运行水位超过警戒水位时，网关进入高负载状态，请求成功率和耗时可能受到影响，建议配置告警。

网关容量指标

节点规格和数量决定网关容量，增大节点规格或者增加节点数量可以提高网关容量。如果节点规格不变，随着节点数量的增长，网关容量基本保持线性增长。网关容量指标说明如下：

- 连接数：网关支持的最大连接数量。
- QPS：网关每秒处理/转发的业务请求总数。
- Token 并发数：网关同时处理的请求中，可支撑的最大 Token 并发生成数。

节点规格		1核2G	2核4G	4核8G	8核16G	16核32G
节点数量		2节点	2节点	2节点	2节点	2节点
连接数	推荐水位	12000	24000	48000	96000	192000
	警戒水位	14400	28800	57600	115200	230400
QPS	推荐水位	1750	2225	7750	10300	125000
	警戒水位	2000	2550	8000	15000	200000
Token 并发数	推荐水位	1000	1000	1000	30000	35000
	警戒水位	2000	2000	2000	50000	50000

网关性能数据

测试场景：客户端发送请求给网关，网关转发到后端服务器，后端服务器直接返回响应。

测试条件：不同的协议类型、连接类型和 Token 分级（Token 分级对应输出 Token 数量，如 L1 约 30tokens，L2 约2000tokens）。

测试结果：客户端请求 QPS 如下所示，基于推荐水位（30%负载）下的达标 QPS。请求平均耗时小于 5 毫秒。
CPU 和内存使用率处于推荐水位(30%)。

CPU 和内存使用率均处于推荐水位（30%）								
协议类型	连接类型	Token	1 核 2G		2 核 4G		4 核 8G	
			2 节点	3节点	2 节点	3节点	2 节点	3节点
HTTP	短连接	30	1600	2500	1500	2500	4300	4800
		2000	1500	2200	1400	2400	3000	2200
	长连接	30	650	900	500	900	900	1400
		2000	600	800	430	800	800	1300
HTTPS	短连接	30	900	1300	800	1100	1350	2400
		2000	900	1000	600	1000	1600	2200
	长连接	30	580	1000	900	880	850	1500
		2000	550	1000	860	700	800	1000

版本生命周期管理

最近更新时间：2026-07-23 17:07:31

为保障用户业务的稳定性和安全性，本文为您介绍云原生网关的版本生命周期管理机制。

版本号规范

云原生网关采用 3 位版本号，即 a.b.c，其中每位版本号含义如下：

- a.b：与开源社区 Kong 的前两位版本号保持一致
- c：需求开发和缺陷修复版本号。如果有需求开发或缺陷修复，增加此版本号。

示例：版本号 2.8.1 表示基于开源 Kong 2.8 版本提供的功能版本。

关键生命周期里程碑

云原生网关为每个发布的产品版本提供长达 27 个月（18 + 9）的服务支持。支持周期分为以下三个阶段：

关键里程碑点	说明	用户需要做什么
GA (General Availability)	指当前版本可全量交付现网客户，在此期间，平台为该阶段中的版本提供补丁和服务支持。	无需额外关注，正常使用产品即可。
EOM (End of Marketing)	指当前版本实例停止全网新建的时间，一般在进入 GA 的18个月后进进入 EOM 阶段。在此期间，平台为该阶段中的版本提供补丁和服务支持。	需要制定升级计划，以便在版本 EOS 前，升级到最新稳定版本。
EOS (End of Service & Support)	指停止为当前版本提供服务的时间，此时版本已过期，一般在进入 GA 的 27 个月后进进入 EOS 阶段。产品将不再提供除版本升级外的其他技术支持，也不再承诺 SLA。	您需要立即采取运维行动，将实例升级到最新稳定版本。在此之前，您的实例可能处于高级别风险之中，同时面临系统被攻击或业务稳定性风险，请务必尽快采取行动。

补丁和服务支持

补丁范围

平台提供的产品新特性、功能缺陷修复、社区特性合并、安全风险修复等。

服务支持范围

- 实例创建：支持创建 GA 阶段的版本实例。
- 升级和运维保障：提供版本的升级功能，并提供问题排障、故障恢复等支持。

- 售后支持：提供答疑、在线指导、排查、排错等工作，但对于过期版本的实例，平台将不保证技术支持的质量和有效性，同时 SLA 可能因未遵守最佳实践而受到影响。

过期版本风险

- 不支持新建过期版本的新实例。
- 平台不再为过期版本提供补丁服务。
- 无法保证技术支持的质量和有效性。
- 平台保留强制升级过期版本实例的权利，执行强制升级操作前，我们会至少提前一个月以短信、邮件、站内信等方式发送相关通知。

为了保证您生产业务的稳定，建议您及时规划并升级实例版本，产品升级操作请参考 [升级网关版本](#)。

按量计费转包年包月

最近更新时间：2026-07-23 17:07:30

场景说明

腾讯云云原生智能网关支持计费模式的转换，提供从按量计费转为包年包月的能力，以满足您灵活地购买和使用引擎。本文档介绍在腾讯微服务平台控制台进行按量计费实例转换成包年包月实例的功能操作。

⚠ 注意：

按量计费转换成包年包月过程中，不会对您的业务访问造成任何影响，请放心操作。

转换规则

- 按量计费转包年包月后，包年包月的引擎实例**无法转回**按量计费的实例，请谨慎确认，以免造成成本浪费。
- 包年包月计费介绍请参见 [计费概述](#)。

操作步骤

- 登录 [腾讯云微服务平台控制台](#)，在左侧导航栏单击 AI 网关或云原生网关，在实例列表找到所需实例，在操作列选择**更多 > 按量计费转包年包月**。

新建 请输入名称或者ID 🔍 🔄									
ID/名称	地域	引擎版本	产品版本	规格	运行/总节点个数	状态	创建时间	计费模式	操作
...	广州	...	标准版	1核/2G	5/5	运行中	...	按量付费	删除 更多
...	广州	...	标准版	4核/8G	5/5	运行中	...	按量付费	重启 节点数变更 节点规格变更 按量计费转包年包月
...	广州	...	标准版	4核/8G	5/5	运行中	...	按量付费	

- 在弹出的对话框，选择续费时长，勾选同意规则，单击**确定**。

💡 说明：

您也可以勾选自动续费，方便您后续的续费管理。

退费说明

最近更新时间：2026-07-23 17:07:30

按量计费退费

按量计费实例可随时释放实例，释放完成后计费将终止。

包年包月退费

说明：

包年包月实例暂不支持到期前自助销毁，如需销毁退费请[提交工单](#)获取协助。

五天内无理由全额退款

退款说明

产品遵守腾讯云 [云服务退货说明](#)，若您在购买注册中心实例后有任何不满意，我们支持五天内无理由退还，具体规则如下：

- 对于每个主体，自新购实例之日起5天内（含5天），可享受1个实例5天无理由退还。退款金额为**购买时花费的全部消耗金额**，包括现金账户金额、收益转入账户金额以及赠送账户金额。
- 后付费的产品切换到**预付费**后不支持五天内无理由全额退款。
- 如出现疑似异常/恶意退货，腾讯云有权拒绝您的退货申请。

退款金额及途径

购买时使用的代金券不支持退还，购买时使用的非代金券费用按支付方式（现金/赠送金）及支付比例退还到支付方腾讯云账户。

非全额退款

不满足五天内无理由全额退款规则的预付费云产品退货退款，退款说明如下：

退款说明

- 切换计费模式并且已退款的资源，发起退款则不重复计算。
- 单价和折扣按照系统当前的优惠为准。
- 参与活动购买的产品，如若退款规则与活动规则冲突，以活动规则为准，活动规则中若说明不支持退款，则无法申请退款。
- 推广奖励渠道订单不支持自助退款，请提交工单发起退款申请。
- 如出现疑似异常/恶意退货，腾讯云有权否决退货申请。

退款金额及途径

退款金额 = 订单实付金额 - 资源已消耗金额

方式基于**按量计费**计算：

- 用户发起退费的当天，已满整月的，按已使用整月的包年包月价格及对应折扣（匹配系统当前折扣）进行扣除；不满整月的，按已使用时长的按量计费价格进行扣除。
- 计算公式：已消耗金额 = 包月价格 × 已使用整月月份 × 当前折扣 + 按量计费时长 × 当前按量单价 × 当前折扣

购买时使用的代金券不支持退还，购买时使用的非代金券费用按支付方式（现金/赠送金）及支付比例退还到支付方腾讯云账户。

欠费说明

最近更新时间：2026-07-23 17:07:30

注意事项

- 引擎实例不再使用时请及时销毁，以免继续扣费。
- 引擎被销毁/回收后，数据将会被清除且不可找回。
- 由于您的实际资源消耗可能不断变化，余额预警可能存在一定的误差。

欠费提醒

当您的账户余额被扣为负值时，我们将通过邮件及短信的方式通知到腾讯云账户的创建者以及全局资源协作者、财务协作者。

欠费处理

从您的账户余额被扣为负值时刻起，在24小时内可继续使用且继续扣费，24小时后将停止服务，不再产生资源占用费用。

停止服务后，系统将对引擎实例进行如下处理：

停止服务后的时间	说明
≤ 7天	若充值至余额大于0，计费将继续，用户可重启引擎。若您的账户余额尚未充值到大于0，则无法重启引擎。
> 7天	若您的账户余额未充值到大于0，引擎资源将被销毁，所有数据将被清理，且不可找回。引擎被销毁时，我们将通过邮件及短信的方式通知到腾讯云账户的创建者以及所有协作者。

 注意

- 请在收到欠费通知后，及时前往控制台 [充值中心](#) 进行充值，以免影响您的业务。
- 如您对消费明细有疑问，可以在控制台 [资源账单](#) 页面查阅和核对您的消费明细。
- 如您对具体的扣费项有疑问，可以参见 AI 网关 [价格说明](#) 及云原生网关 [价格说明](#) 了解具体的计费项含义及计费规则。
- 您还可以通过计费中心的余额告警功能，自行设定欠费预警。详细信息请参见 [余额预警指引](#)。