目标:

- 1、阿里云的用户快速部署 NVIDIA AIE, 默认支持 90 天的 POC 测试。
- 2、阿里云的用户在云市场采购 NVIDIA AIE 的 License, 用阿里云的余额或信控支付。

前置条件:

- 1、已经完成阿里云账号的注册并完成实名认证。
- 2、阿里云的账户余额或者信控额度足够支付 License 的费用。

## 目标1步骤:

- 1、申请 NVIDIA NGC API Key
- 入口: <u>https://build.nvidia.com/meta/llama3-8b</u>
- 步骤: Experience tab -> 页面右侧的 Docker ->填写邮箱-> Get Api Key -> Self-Hosted
- API , 然后就生成了下图所示的 API Key
- 用途:用于从 NGC 拉取 NIM docker image,在部署中也是一个必填的参数

Integrate NIM into your application Use a hosted API or self-host on your own infrastructure Hosted API Self-Hosted API	Even for the form	
Step 1 of 2	Contraction of the second second second	
Generate your API Key Create a new API key if you don't already have one. You will need this in the next step.		ence
nvapi-t8xsfV9Vp6NZ9-AcFHFSEMSy6YqXrQowt8eYeug4MbMua5EH-tzGı	The balance contransists will devertised as a 2-3-3-the exploring of VVDR A for installation 2-302-instruct on your GPU. It is using the large exploring explored as a 2-3-3-3-3-3-3-3-3-3-3-3-3-3-3-3-3-3-3-3	ence
$\downarrow$ — — – – – – – – – – – – – – – – – – –	E a mini fragm anno de Branne al manare Manare de Martine et al portado tra en adorado mais for para sampla manament.	×
Step 2 of 2	E a super NULLER (AF-100000 Hill of All Alls) E a super Station (All All All All All All Alls) E avece - a "All R. All(AllS) E Autor - cal - and - N quark (All - quark (All -	
Pull and run the NIM locally Click the Docker tab and follow the instructions to pull and run the NIM locally.		
★ You have access to exclusive NVIDIA AI Enterprise features		
View <u>software security &amp; vulnerability reports</u> & get <u>enterprise grade support</u> .		
Got it!		

- 2、登陆阿里云账号
- 3、访问页面: https://market.aliyun.com/products/201214006/cmgj00067281.html
- 4、点击页面「立即开通」按钮,并进行留资。

	基于NVIDIA NIM快速部署LLM模型推理服务										
$\checkmark$	NVIDIA NIM 是 NVIDIA AI Enterprise 的一部分, 是一套易于使用的预构建容器工具, 目的是帮助企业加速生成式 AI 的部署。它支持各种 AI 模型,可确保利用行业标准 API 在 本地或云端进行无缝、可扩展的 AI 推理。										
	近180天成交:小于10单 评论: 0 ★ (0条)										
	交付方式;计算巢私有化部署										
	<b>免费开通</b> 按使用量计 <b>费查看价格</b>										
	<u>立即开通</u>										

勾选「同意《服务商用户协议》与《云市场平台服务协议》」并留资



5、创建服务实例,根据页面显示的内容配置环境信息

该服务会根据配置信息在当前登录的阿里云账号下创建 NVIDIA AIE 在阿里云的系统环境, 并启动执行。

启动后的环境自带 90 天 POC 测试用的 License, 但是运行环境所需要的阿里云 ECS 以及 相关的资源的费用是通过当前部署服务的阿里云账号向阿里云支付,软件本身不会产生额外 的费用。

创建服务实例

● 基于NVIDIA №	NIM快速部署LLM模型推理服务-计算巢社区服务	有名称 上海曙云信息技术有限公司
	下 .	业 │ ⇒ ☴ 必填 1/4 全部 9/14
服务实例名称	请输入服务实例名称 	服务实例名称 * 地域 ✓ • 软件付费配置 付费方式
* 地域	华东1 (杭州) •	<ul> <li>NIVIDA NGC 配置</li> <li>* NVIDIA NGC API Key ×</li> <li>付费类型配置</li> </ul>
<b>软件付费配置</b> 〈	按量付费	<ul> <li>Kubernetes記置</li> <li>* Worker节点规格 ×</li> <li>Worker 系統母磁母光則</li> </ul>
		Worker节点系统盘大小(GB) Service CIDR Pod 网络 CIDR
NIVIDA NGC 配置 * NVIDIA NGC API Key	@	<ul> <li>实例密码</li> <li>可用区配置</li> <li>★ 可用区</li> </ul>
	API Key 是一个身份验证令牌,从NVIDIA的NGC仓库下载模型和容器需要使用API Key进行身份验证,参考文档: https://docs.nvidia.com/nim/large-language-models/latest/getting-started.html#generate-an-api-key	- 网络配置 专有网络IPv4网段

6、下一步进入付款步骤, 请确保账户余额可支付阿里云云资源的采购费用

<b>软件费用</b> 软件费用:¥0.0 质价:¥0.0 0	℃.0				
<b>阿里云产品费用</b> 预估算费用仅供	参考,最终金额以实际产生的费	用为准。			导出文件
g时长费用:¥ <b>30.409</b> /小时	CU容量数量: 1 个 ⑦ 按流	z量数量: <mark>1</mark> 个 ⑦			
资源类型	资源用途	规格	地域	数量	预估算费用
	ManagedKubernetesCluster	-	华东 1 (杭州)	1	¥0.64/小时
	EcsInstanceWorker	实例: ecs.gn7i-c8g1.2xlarge (GPU 计算型 gn7i, 8 vCPU, 30 GiB) 系统盘: cloud_essd 120GiB	华东 1 (杭州)	3	ecs.gn7i-c8g1.2xlarge:¥29.354/小时
	SIbInstanceIngress	slb.s1.small (简约型I)	华东 1 (杭州)	1	¥0.1/小时
Kubernetes托管版集群 (CS)	SIbInstanceApiServer	slb.s1.small (简约型I)	华东 1 (杭州)	1	¥0.1/小时
	NatGateway	-	华东 1 (杭州)	1	¥0.195/小时 CU容量费用: ¥0.195/CU
	Eip	-	华东 1 (杭州)	1	¥0.02/小时 流量费用: ¥0.8/GB
文件系统(NAS)	NasFileSystem	-	华东 1 (杭州)	1	¥0/小时
您当前账户余额为 ¥ 97.24, 可用额	度为 ¥97.24,信控余额为 ¥0	00			

勾选服务条款后,就可以进入支付步骤,这里需要注意的是,因为会开通按量付费的资源, 请保证账户余额大于 100 元,否则系统会提示无法满足创建条件。

💶 检测到当前的实例创建存在以下问题

- 11 个依赖项异常。
- 余额不足。您当前账户余额为¥97.24,可用额度为¥97.24,信控余额为 ¥0.00

购买规则:客户的账户余额或代金券、信控额度的综合在 **100元** 以上,才可购买按量付费资源。

检测到当前您的余额不足,建议您充值后再进行部署,否则可能影响后续服 务使用。

可能会造成资源创建失败,是否返回修复后再进行立即支付。

返回修复

7、付款成功后, 阿里云的计算巢系统就会启动部署任务, 大概 10 分钟左右, 系统部署完成, 登陆计算巢可以查看服务的部署信息。

计算巢的控制台地址: <u>https://computenest.console.aliyun.com/</u> 部署好的系统会自带 90 的测试 License。

关于 NVIDIA AIE 的使用, 请查阅 NVIDIA AIE 的相关技术文档: https://docs.nvidia.com/nim/index.html

## 目标2步骤

在进入目标 2 之前,请与客户就 License 的采购事宜达成共识。

阿里云云市场的商品支付客户的阿里云账号余额和信控支付,License 的采购账号与目标 1 的阿里云账号可以不是同一个账号。

客户下单采购的发票由「上海曙云信息技术有限公司」提供。

关于技术和售前咨询,可以点击商品右上方的「NVIDIA 服务咨询」在线咨询。

		服务利	市 上海塘云信息技术有限公司
	【NVIDIA AIE-License授权】基于NVIDIA NIM快速部署LLM模型推理服务	服务的	能力等级 AI与大数据计算 〇〇〇〇〇〇〇〇
~	NVIDIA AI Enterprise 是一个错到端云原生软件平台,可加速数据科学工作流,简化生产级协作驾驶和其他生成式 AI 应用的开发和部署。易于使用的微磁务优化了模型性能,可提 供企业级的安全性、支持服务和稳定性,能够确保以 AI 为基础开展业务的企业从原型到生产的平稳过度。	£1£1	<ul> <li>技术服务咨询</li> <li>NVIDIA服务咨询</li> </ul>
	近180天成交:小于10单 评论: 0 ★ (0条)		
	交付方式:人工服务	电话	4009030002 转16002
	喜餐板本: POC测试		19921132513
	购买时长: 「揮	ED HA	yong.wang@shuyuninfo.com
	0.01		
	*0.01		♡ 服务商手屋服务
	续费: ¥0.01		◇加労回ミ病服労
		1	您好,服务商可根据您的需求提供1v1咨询服务及方案报
	<u>立即時天</u> 加入购物车	1	们,能更高效地贴近您的需求。
			70021 (4470

【NVIDIA AIE-License 授权】基于 NVIDIA NIM 快速部署 LLM 模型推理服务商品地址: https://market.aliyun.com/products/201214006/cmfw00067311.html

需要注意的是:该商品页面只用于前台信息展示,具体的报价信息会通过云市场的「Private Offer」向客户进行报价,步骤如下:

1、请与客户线下沟通关于 License 的采购事宜,并获取客户的「阿里云账号」或者「阿里 云 UID」,请务必确认信息准确,以确保报价单推送成功。

2、请提供 License 的报价、单位,支付金额是人民币,以上信息请邮件发送给郑飞(邮箱: philipsz@nvidia.com),客户通过阿里云账号付款,云市场的上架服务费请与郑飞沟通。

确定报价

Х

发送给客户的报价单将包括您选配的商品描述、服务详情与报价。 商品价格: 0.01 元 商品报价: 1 元



当完成「Private Offer」的推送后,用户的阿里云账号会收到短信和邮件提醒

短信 今天 11:02

【阿里云】尊敬的用户:阿里云市场已为您提交定制方案,请登录阿里云控制台 - 云市场 - 定制方案管理 查看。

(一)阿里云	产品	解决方案	文档与社区	免费试用	定价	云市场	了解阿里云
茵為的田户·							
服务商已为您在云市场提交的需求发送	了报价单,请点	击方案详情查看。					
温馨提示: 请仔细查看方案价格与服务	条款是否同服务	务商与您约定的一致	致。				

或者引导客户进入「Private Offer 」页面查看报价单信息: https://market.console.aliyun.com/#/privateoffer

三(一)阿里云	☆ 工作台				Q 證素	费用 ICP	备案 企业 支持 工	# ⊑ 4 <sup>°</sup> ®	🚷 簡体	ni318*****@— 🥎 主账号 🦤
云市场	PrivateOffer管理									
已购买的服务	方案ID	方案名称	服务商	关联需求	提交时间	方案总报价	方案状态			操作
订单列表	1725591756623	NVIDIA AIE的报价单	上海曙云信息技术有限 公司 (产品服务商)		2024-09-06 11:02:37	1	未支付			宣看详情
来支付订单 需求管理	1714034592063	人脸实人认证5折忧患	山东蓝笛信息科技有限 公司 (产品服务商)		2024-04-25 16:42:02	9	未支付			查看详情
PrivateOffer管理	1709782398071	三月采购季1分钱体验【低至0.13元】	山东蓝笛信息科技有限 公司 (产品服务商)		2024-03-07 11:32:11	0.01	未支付			查看详情
追款管理 发票管理								共有3条,每页量	示: 10条 🔍	< 1 > Ø

勾选代表您已阅读并同意《云市场平台服务协议》《【NVIDIA AIE-License 授权】基于 NVIDIA NIM 快速部署 LLM 模型推理服务-商品在线协议》即可进行支付。

PrivateOffer管理 全返回上	—级						
方案IC	: 1725591756623						
方案名称	: NVIDIA AIE的报价单						
服务商	: 上海曙云信息技术有降	限公司 (产品服务商)					
关联需求	:						
方案添加时间	: 2024-09-06 11:02:37						
方案总报价	: 1						
方案状态	: 未支付						
方案报价单	商品code	商品名称	商品价格	配置详情	购买周期	数量	支付状态
	cmfw00067311	【NVIDIA AIE-License授权】 基于NVIDIA NIM快速部署 LLM模型推理服务	¥1	套餐版本:POC测试	1年	1	未支付
方案备注	这付						
	🛃 勾选代表您已阅读	载并同意《云市场平台服务协议》《 <b>【</b>	NVIDIA AIE-Li	cense授权】基于NVIDIA NIM	1快速部署LLM模型推理	服务-商品在线协议》	

C→ 阿里云	冒 购物车 1	工单 备	纂 简体中文 ▼	hi3185****@aliyun.c
支付				
云帝场订单变付方式已升级,选择赢付遇/融景或课道法行变付,资金有重接付款到局品的高家账户。				
待支付订单				应付:¥1.00 ×
び#号 产品 配置 数量		时长		总应付金额
INVOIDA (FL:Consertient)         第           2381/25331150642         予約100A NMU快速率以LM模 第二次         第二次         1           1         加速運動         第二次         1         1		1年		¥1.00
使用余额				抵扣:¥1.00
遗得支付方式				支付: ¥0.00 <

当客户完成付款后, 服务商会将 License 通过阿里云云市场进行交付, 服务商提交了交付信息后, 客户的阿里云账号会收到邮件和短信提醒:

【阿里云】 尊敬的用户:您购买的【NVIDIA AIE-License 授权】 基于 NVIDIA NIM 快速部署 LLM 模型推理服务产品已开 始生产,请登录阿里云 "管理控制台-云市场-已购买的服务" 关注生产进展。

【阿里云】尊敬的用户: 您购买的【NVIDIA AIE-License 授权】基于NVIDIA NIM快速部署LLM模型推理服务产品,已 经全部为您服务完成,请确认是否满足要求,系统将在1.00个工作日后自动确认,并关闭该服务。您可以登陆控制 台进行确认: <u>https://t.aliyun.com/4QCRrzCf</u>

<b>(一)</b> 阿里云	产品	解决方案	文档与社区	免费试用	定价	云市场	了解阿里云
----------------	----	------	-------	------	----	-----	-------

尊敬的hi31854864@aliyun.com:

您购买的云市场产品【NVIDIA AIE-License授权】基于NVIDIA NIM快速部署LLM模型推理服务已经开始生产。

您可以登录阿里云"管理控制台-云市场-已购买的服务"关注生产进展。点此进入

## 客户登陆阿里云账号在云市场的交付中心收到交付的 License 信息:

https://market.console.aliyun.com/imageconsole/index.htm

三()阿里云	の 工作台		QB	表示	I用 ICP 备案 企业 支持 工单	🖬 🗘 🕐 🚫 📾 🕨	ii318*****@ 🧐
云市场	巴购买的服务						
已购买的服务		组合套餐,一起买	,更划算!更方便!				
订单列表 未支付订单	商品名称 ~ 请输入				=,,		
需求管理	商品	商品类型 🏆	状态(已遗6项状态) 🌱	到期时间 🍸	是否试用实例 🆓		操作
PrivateOffer管理 退款管理	【NVIDIA AIE-License授权】基于NVIDIA NIM快速 部署LLM模型推理服务 设置备注 /	服务	服务中	2025-09-07	非试用实例		
发票管理	规格: POC测试 付费方式: 闹期购买						2
优惠券管理 合同管理	创建时间:2024-09-06 实例ID:94191623	上海塘云信思技不有限公司	₩ 喉水弦()				
我的关注					<b>流程当前</b> 项目交付评	存点:提供授权序列号 处理角色:服务者 情可点由交付中心重看	( <u>交付中心</u> 把

C-D	交付中心								
۵	项目详情								文件管理 项目留言 流程日志
	第1步	0	⑦ 分配服务商节点				2024-09-06 11:08:44	E	上海曠云信息技术有限公司 交付商管理员
	第2步	0	③ 分单	查看上传内容			× 14	a	D 13501684497
	第3步	0	<ul> <li>         ・ 損供授权序列号         ・ 2024-09-06 11/08-44 上海場正成意法术有限に         ・ 2024-09-06 11/15-30 上海場正成意法术有限に         上传描述: ABCDEFGHIZKLMN         </li> </ul>	ABCDEFGHIZKLMN 文件名称	上传时间	大小	i4 操作	业务信息 <ul> <li>∴产品名称:【M</li> <li>服务</li> <li>产品规格: POI</li> <li>①単号: 2381</li> <li>寄户名称: N3</li> <li>到期时间: 202</li> </ul>	VIDIA AIE-License現在] 基于NVIDIA NM快速搭載LLM模型推奨 CMIE 285551150662 1854864@shjun.com 55-09-07
	<b>第</b> 4步	0	用户完成确认 • 3024-09-06 11:15:30 系统 进入用户完成确认		留无数据				

客户验收完 License 后,请务必让客户点「确认验收」,并给与5 星好评,感谢支持。



## 当客户完成了验收后,可以申请开票 进入「费用-发票」页面,勾选对应的订单:

					<b>—</b> • • • • • • • • • • • • • • • • • • •												
≡	<b>(-)</b> 阿里云	6	工作台						Q 搜索		费用 ICP 备案	企业 支持 工单 🖂	4 🕄 🚷 🍈	hi318*****@. 主账4			
费用与	可成本	Q		金造()約页):共44个订单可申请发票 总金额:¥1,649.79													
				۰.	订单/账单/合同编号	开票主体	单振类型	账单类型	产品名称	应付现金金额	可开票金额	关联订单/账单/合同编号	资源实例ID	28			
皆页					238125351150662	上海曙云信息技术有限公司	订单	新购	【NVIDIA AIE-License授权	¥ 1.00	¥ 1.00	-	94191623	20			
98/**		Ť			237951356520662	阿里云服务	订单	新购	无影云电脑时长包	¥ 98.99	¥ 98.99	-	tp-0bizsc1aoyfhhd338	20			
市田					237565777360662	阿里云服务	订单	新购	个人版云电脑月套包	¥ 14.90	¥ 14.90	-	pmdp-0brgb5n0xdpkgfe	cu 20			
TRO	тю	~			202408	阿里云服务	账单	新购	后付费结算单-202408-快照	₩ 7.44	¥ 7.44	-		20			
10 M3					202408	阿里云服务	账单	新购	后付费结算单-202408-云	¥ 4.51	¥ 4.51	-	-	20			
**					202407	阿里云服务	账单	新购	后付费结算单-202407-快照	¥ 7.44	¥ 7.44	-	-	20			
发票					202407	阿里云服务	账单	新购	后付费结算单-202407-云	¥ 0.30	¥ 0.30	-	-	20			

账号信息	发票预览 ④					
资源购买账号: hi31854864@aliyun.com	电子发票(增	值税专用发票) x #9%:				
关键信息		58				
发票抬头: 个人	至一名称: 个人 行 统一社会信用代码(纳税人识别号: 夏	各称: 3000000000000000000000000000000000000				
发票类型: 数字化电子专票 (电子) シ	货物或应税劳务、服务名称 规格型号	单位 說服 全朝 校準 秋期				
接收邮箱: zhangyu2590@163.com 选择邮箱	\$2件描绘费 余 11	\$ 1 ¥0.94 6.00% ¥0.06 ¥0.96 ¥0.9				
基本信息	(外稿会计 (大写) 夏元間	(i)(lij) ¥ 1.00				
开票内容 规格型号 数量 税率 金額	Ē					
软件服务费 空白(发票上空白, 不显示 V 1 套 6% 1	开观人:					
发票备注:(建议不要超过25个汉字成50个(数字+字母),否则盖章的时候有可能压到。)						
该发票由服务商提供,请联系服务商确认将备注添加到发票上。						

服务商收到开票申请后, 会在2个工作日内根据提交的申请信息完成开票。