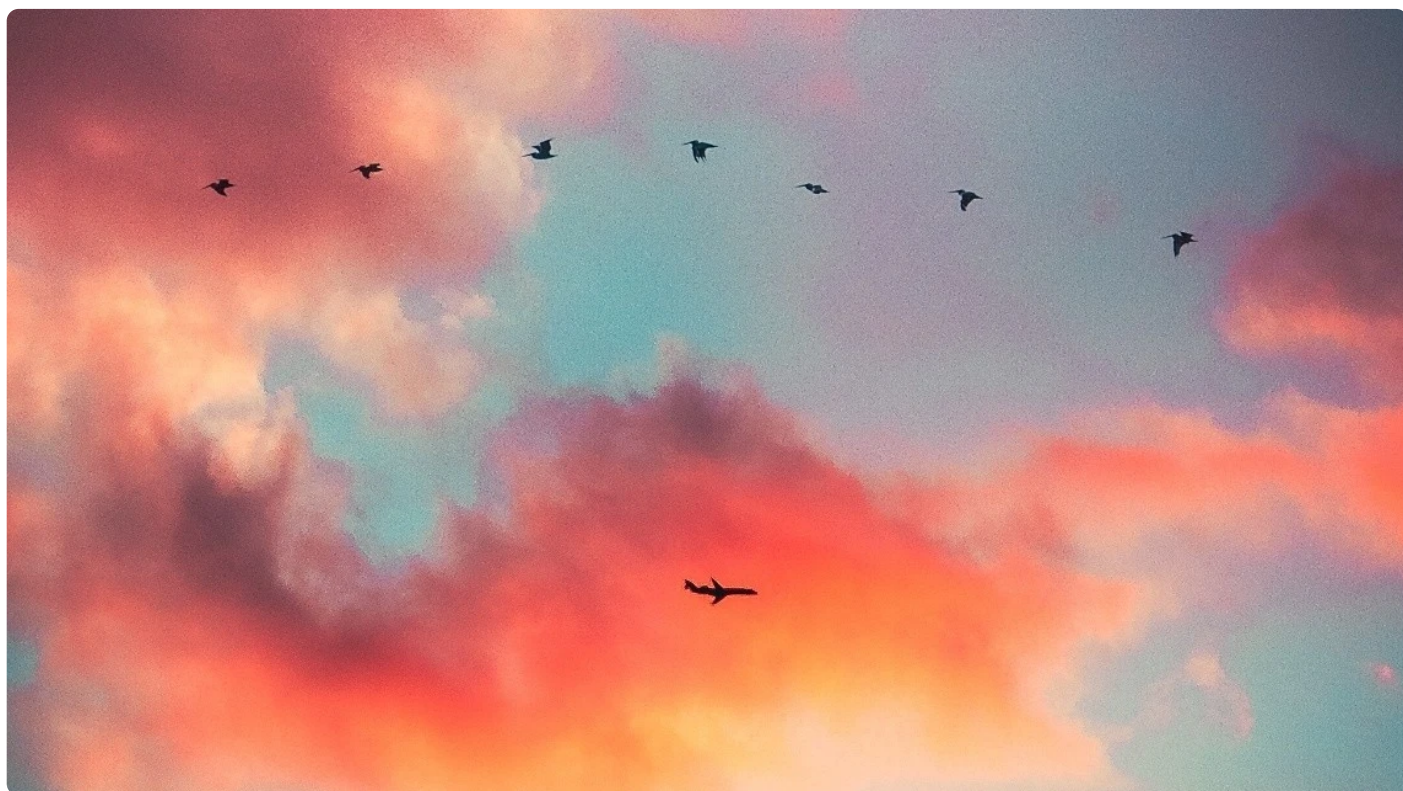


03 | 云虚拟机（二）：眼花缭乱的虚拟机型号，我该如何选择？

2020-03-09 何恺铎 来自北京

《深入浅出云计算》



你好，我是何恺铎。

在上一讲中，我带你了解了云虚拟机的大致构架和组成，实际体验了在云上建立第一台虚拟服务器的完整流程，还介绍了在创建过程中，你所需要注意的若干重要选项及其含义。

而在这些选项之中，最重要的恐怕就是**虚拟机的规格**了，因为它直接决定了虚拟机的计算能力和特点，同时，也会深刻地影响使用成本，是你在选型时需要考虑的重点问题。

很多同学在实际工作中，都会遇到这样的困惑：公司要上云，或者因为业务发展需要采购新的云服务器，但是在查看某云厂商的官网时，发现可选择的虚拟机型号列表很长，有点儿眼花缭乱。

那么，不同种类的虚拟机到底有什么区别呢？在选择时又应该从哪儿入手呢？

今天，我们就来详细聊聊这个话题。

建立对虚拟机配置的多维认知

完整形容一个虚拟机的核心配置和能力，需要从多个角度来入手和描述。弄懂了这些重要维度的含义，你才能够准确理解一个虚拟机的性能预期和使用场景，从而作出正确的型号选择。这里并非只有决定 CPU 核数和内存大小这么简单。那么，主要是哪几个维度呢？

第一个维度，就是虚拟机的“类型”，或者说“系列”。

这是一个非常重要的概念，它是指具有同一类设计目的或性能特点的虚拟机类别。

一般来讲，云厂商会提供通用均衡型、计算密集型、内存优化型、图形计算型等常见的虚拟机类型。这些类型对应着硬件资源的某种合理配比或针对性强化，方便你在面向不同场景时，选择最合适的那个型号。

而 vCPU 数和内存大小（按 GB 计算）的比例，是决定和区分虚拟机类型的重要指征之一。

通用均衡型的比例通常是 1:4，如 2 核 8G，这是一个经典的搭配，可用于建站、应用服务等各种常见负载，比如作为官网和企业应用程序的后端服务器等。如果你对未来工作负载的特征还没有经验和把握，那你也可以先使用通用型实例，等程序运行一段时间后再根据资源占用情况按需调整。

如果 vCPU 和内存比是 1:2 甚至 1:1，那就是**计算密集型**的范畴，它可以用于进行科学计算、视频编码、代码编译等计算密集型负载。

比例为 1:8 及以上，一般就会被归入**内存优化型**了，比如 8 核 64G 的搭配，它在数据库、缓存服务、大数据分析等应用场景较为常见。

图形计算型很好理解，就是带有 GPU 能力的虚拟机，一般用于机器学习和深度学习模型的训练和推理。随着 AI 的火热，这类机器也越来越多地出现在各种研发和生产环境中。

在主流云计算平台上，常常使用字母缩写来表达虚拟机系列。比如，AWS 的通用型是 M 系列，阿里云的内存优化型为 R 系列，Azure 的计算优化型为 F 系列等。

不同云平台之间使用的字母可能相同，也可能大相径庭，你在记忆时需要小心，不要混淆。在这里，我根据各家 2020 年的最新情况，简单整理了一个表格供你参考：

虚拟机类型/系列	AWS字母代号	阿里云字母代号	Azure字母代号
通用均衡型	M	G	D, A
计算密集型	C	C	F
内存优化型	R	R	E
图形计算型	P	GN	NC, ND, NV
本地存储型	I, D	I, D	Ls

需要注意的是，上表中还提到了本地存储型，它是指带有高性能或大容量的本地存储的机型。我们在后续讨论云盘的课程中还会提到，这里你先了解一下就可以了。

第二个重要的维度，是虚拟机的“代” (Generation)，用来标识这是该系列下第几代的机型。

我们知道，数据中心硬件和虚拟化技术是在不断发展的，云厂商需要不断地将最新的技术和能力推向市场，让你享受到时代进步带来的技术提升。这和我们个人用的笔记本电脑是非常类似的，笔记本厂商也总是在不断地更新设计和配置，以赢得消费者的青睐。**所以即便是同一系列的机型，不同的代别之间也会有不小的区别。**

具体来讲呢，同类型虚拟机的更新换代，往往首先会带来相应硬件 CPU 的换代提升。随着一代新机型的推出，云厂商一般都会详细说明背后支撑的硬件详细信息。

比如说，AWS 在 2017 年末，在全球发布的新一代 EC2 实例 M5/C5/R5，它们的背后是升级到了 Skylake 架构的 Intel 至强铂金系列处理器，相比前一代采用的 Broadwell 或 Haswell

架构处理器，进步了不小，还支持了可大幅提升矢量和浮点运算能力的 AVX-512 指令集。

再比如，阿里云在 2019 年的云栖大会上，也盛大发布了第六代 ECS，它全线采用了更新一代的 Intel 至强 Cascade Lake 处理器，相较前一代的 Skylake 实例，又在性能、价格优势等各方面有了进一步提升。你可以参考下面给出的截图：

架构

x86 计算

异构计算 GPU / FPGA / NPU

弹性裸金属服务器 (神龙)

超级计算集群

分类

通用型

计算型

内存型

大数据型

本地 SSD

高主频型

入门级(共享)

规格族	实例规格	vCPU	内存	处理器型号
<div><</div>				
<div><div><input checked="" type="radio"/></div>内存型 r6</div>	ecs.r6.large	2 vCPU	16 GiB	Intel Xeon(Cascade Lake) Platinum 8269CY
<div><div><input type="radio"/></div>内存型 r6</div>	ecs.r6.xlarge	4 vCPU	32 GiB	Intel Xeon(Cascade Lake) Platinum 8269CY
<div><div><input type="radio"/></div>内存型 r6</div>	ecs.r6.2xlarge	8 vCPU	64 GiB	Intel Xeon(Cascade Lake) Platinum 8269CY

阿里云第六代 ECS（内存型）型号选择界面

这里你需要特别注意，正是由于虚拟机所采用的物理 CPU 在不断更新，所以**云上虚拟机的单核性能未必相同**。有时，虽然两个虚拟机的核数一致，但由于底层芯片的架构和频率原因，性能上可能有较大的差别，我们需要注意在不同机型间做好比较和区分。

像微软 Azure，就引入了 Azure Compute Unit (ACU) 的概念，来帮助量化不同 CPU 的单核性能。比如其历史较久的通用型 A 系列，它的单核性能基准为 100 单位，而计算型的 F 系列的单核算力则高达 210~250，是 A 系列的两倍还多。

另外，你还应当看到，**云虚拟机的换代更新并不仅仅只在 CPU 等硬件配置层面，很多时候也伴随着底层软硬件架构的更新和提升，尤其是虚拟化技术的改进。**

前面我提到的 AWS 第 5 代 EC2 实例，正是全面地构建在 AWS 引以为傲的 Nitro System 新一代虚拟化技术栈之上。

Nitro System 的本质，是将许多原来占用宿主机资源的虚拟化管理工作进行了剥离，并将这部分工作负载，通过 Nitro Card 这样的专用硬件进行了硬件化，达到了最大化计算资源利用率的效果。在这一点上，阿里云的神龙架构也采用了类似的做法，与 AWS Nitro 可谓一时瑜亮，有异曲同工之妙。

总的来说，我们消费电子产品时的“买新不买旧”，在云端同样适用。新一代的型号，往往对应着全新的特制底层物理服务器和虚拟化设施，能够给我们提供更高的性能价格比。

所以，有些云平台在选择虚拟机型号时，会贴心地默认隐藏相对过时的型号。当然在个别情况下，比如数据中心的新机型容量不足，或者老型号有促销活动时，你也可以酌情选用之前的型号。

第三个重要的维度，就到了我们所熟知的实例大小（Size），也就是硬件计算资源的规模。

在选定的机器类型和代别下，我们能够自由选择不同的实例大小，以应对不同的计算负载。

如果你只是个人用来实验，那么也许单核或者双核的机器就足够了；如果是要放在大规模的生产环境当中，则可以按需选取高得多的配置，现代云计算已经能够提供多达 128vCPU 的机型了。

在描述实例大小时，业界常常使用 medium、large、xlarge 等字眼来进行命名区分，这样的描述基本已经成为事实标准，包括 AWS、阿里云、腾讯云在内的多家主流厂商都在使用。

我们可以大致这样记忆：**标准 large 对应的是 2vCPU 的配备，xlarge 则代表 4 个 vCPU，而更高的配置一般用 nxlarge 来表达，其中 n 与 xlarge 代表的 4vCPU 是乘法关系。**比如，8xlarge 就说明这是一台 $8 \times 4 = 32$ vCPU 的机器。

注意：这里在进入更严谨的配置表达时，我们更多倾向于使用 vCPU 而非核数（Core）来描述虚拟机处理器的数量。因为超线程（HyperThreading）技术的普遍存在，常常一个核心能够虚拟出两个 vCPU 的算力，但也有些处理器不支持超线程，所以 vCPU 是更合适的表达方式，不容易引起混淆和误解。

在某些场景下，你可能还会看到“metal”或者“bare metal”这样的描述规格的字眼，中文称为“裸金属”。它们就是云服务商尽最大可能将物理裸机以云产品方式暴露出来的实例，主要用于一些追求极致性能，或是需要在非虚拟化环境下运行软件的场景。

理解虚拟机命名规则

经过前面的介绍，我们已经了解了决定虚拟机配置的最重要的三个要素，即**类型**、**代别**和**实例大小**。这样，一个完整的虚拟机型号命名就已经呼之欲出了。我们来看最具代表性的 AWS 命名规则（阿里云采用的也是非常类似的格式）：

 [类型名][代别][后缀(可选)].[规格]

这其实就是利用上述的各维度，按照某种顺序排列的一个组合。理解了这一点，当你今后看到某个具体型号的时候，就能够很快地明白该型号命名背后的含义了。

比如，对于 **r5.4xlarge** 这个型号，我们会很快想到，这首先是一个 R 类型的第 5 代的内存型机器，它应该有 $4 \times 4 = 16$ 个 vCPU，内存大小则是 $16 \times 8 = 128\text{G}$ （内存型机器的 CPU 内存比一般为 1:8）。这样分解下来，原来看上去比较陌生晦涩的一个字符串，是不是就立刻变得清晰起来了？

当然，并非所有的云都一定是采用类似 AWS 的命名规则，像是微软 Azure，就用了一个略有不同的命名体系，大致可以总结为：

 [类型名][规格(数字表达)]v[代别]

比如“**E4 v3**”，就代表了微软 Azure 上 4 核 32G 的第三代内存型机器。掌握了 Azure 的格式特征后，你同样能够很快地解读标识的具体含义。


不知道你有没有注意到，在前面的命名公式中，还有一个我们称之为“**后缀**”的可选部分，在许多的型号命名中都能看到它。这个可选部分呢，它一般是作为型号硬件信息的一个重要补充，这种型号与不带此后缀的标准版本相比，会有一些显著的区别或特点，这也是你需要重点关注的地方。

这里我给你举一些型号后缀的例子吧。

比如，AMD 现在凭借 EPYC 霄龙芯片，也开始在服务器硬件市场攻城拔寨，许多云厂商就专门推出了使用 AMD CPU 的云虚拟机，这些虚拟机往往会使用字母 a 作为后缀。AWS 上的 **m5a 型号**，就是使用 AMD EPYC 7000 系列服务器 CPU 构建的通用型虚拟机。

再比如，AWS 的 C5n 计算型虚拟机，其中“n”这个后缀表达的是，该规格在网络层面进行了增强，会比同型号标准机型拥有更大的带宽和网络吞吐能力。在阿里云上，表达相同“网络增强”含义的后缀则是“ne”。

有时，为了验证机型配置是否与我们的期望相符，在 Linux 环境下，我们可以使用 **lscpu** 命令，来了解手中虚拟机的 CPU 信息，并与机器的具体型号名称进行对照。下面的信息，是我在一台 AWS 的 **m5a.xlarge 机型**上运行的结果，你可以看到芯片提供商 AMD 及双核四线程等关键信息，与机型命名的含义相符：

 复制代码

```
1 [ec2-user@ip-xx-yy-zz ~]$ lscpu
2 Architecture:          x86_64
3 CPU op-mode(s):        32-bit, 64-bit
4 Byte Order:            Little Endian
5 CPU(s):                 4
6 On-line CPU(s) list:   0-3
7 Thread(s) per core:    2
8 Core(s) per socket:    2
9 Socket(s):              1
10 NUMA node(s):          1
11 Vendor ID:              AuthenticAMD
12 CPU family:             23
13
```

```
14 Model: 1
15 Model name: AMD EPYC 7571
16 Stepping: 2
17 CPU MHz: 2379.224
18 BogomIPS: 4399.39
19 Hypervisor vendor: KVM
20 Virtualization type: full
```

课堂总结与思考

今天，我们主要探讨了云上虚拟机的类型与规格，相关要点可总结如下：

云虚拟机的配置规格主要取决于类型、代别、实例大小三个最重要的维度。

实例所属的类型，决定了虚拟机相应的硬件资源配比与专项能力，分别为不同的场景优化设计。你可以根据实际场景来酌情选用，这样既能满足需求又好控制成本。

云虚拟机的型号名称一般由类型、代别、实例大小这几项的缩写组合而成，有时还会带有补充后缀。了解了某个云的型号格式后，通过拆分对应，你很容易理解具体型号的含义。

最后，作为今天的交流讨论题，你可以回忆一下，在生产或测试环境中，使用过的最强劲的云端机型。你注意过它是什么系列、什么型号的吗？它主要被用于什么业务场景呢？

欢迎在留言区和我互动，我会第一时间给你反馈。如果觉得有收获，也欢迎你把这篇文章分享给你的朋友。我是何恺铎，感谢阅读，我们下期再见。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

精选留言 (19)



何恺铎 置顶

2020-04-18

[上讲问题参考回答]

1. 一般常见的方法是通过“跳板机”来间接地访问“内网”的机器，跳板机可以通过使用“弹性IP”来向公网开放。另外，还可以使用VPN来让你的客户端连接到内网虚拟机。相关概念在第6讲

中会有进一步讲解。

2. 判断“关机”收不收费的关键，要看相关的资源是不是被“释放”了。在大多数云上，默认的关机将虚拟机的CPU和内存资源彻底释放，但云硬盘一般会为你保留，所以这时虚拟机不收费，但硬盘可能收取少量费用。还需要注意，部分云中有另一种关机的模式（比如阿里云的“停机并继续收费”和Azure的操作系统内关机），这时CPU和内存资源其实并未释放，所以会继续收费，但这种关机模式能够避免后续因资源不足而开机失败和实例开机后漂移到另一台宿主机。



15



cloudwas 置顶

2020-03-09

1. 申请数据库类型的虚机，cpu和内存可能会大一些16C32G，
2. 对于虚机名的生成，我们平台在设计时，为了更好的满足不同的客户需求，主机名是客户自定义规则，比如 [虚机名前缀][用途] [负责人] [ip] [index][规格]等

作者回复: 主机名的确需要妥善命名，不然很难维护。当然，你也可以考虑利用云上的资源标签(tag)功能，来记录你上面提到的部分属性信息，这样能够更容易地筛选过滤。



7



拉斯 置顶

2020-03-10

老师还少讲了一个内网网络类型，从10G，到25G，甚至100G，对于那些AI数据训练很关键。

作者回复: 内网网络带宽，的确是虚拟机配置的一部分，谢谢你的补充。这里我们没有提，其实是因为留在第6讲中作为了讨论题。



6



我来也 置顶

2020-03-09

到目前为止,用过最奢侈的云端机型,也就是阿里云的ecs.g6.xlarge. 4CPU 16G内存了. 平常都是几个阿里云ecs.t5-c1m2.xlarge 4CPU8G内存的.

上家公司用的最豪的配置就是4CPU8G内存的配置了. 生成环境才一台这种配置,要支撑同时过万人的服务.

由于是golang开发的服务,平常的内存使用量比较低.

不像java的,开一个kafka的docker就是1G的实际占用内存.开一个es,又是一个多G的实际占用内存.

作者回复: 赞。用最低的配置, 干最多的事情, 是程序员的不懈追求~

共 2 条评论 >

👍 5



旺旺 置顶

2020-03-10

我想问在选择机器的时候, 都跟CPU的主频没有关系吗?
还是说都不用看主频的, 或者没法看主频呢?

作者回复: 有关系的, 比如不同代的机器, CPU因为型号不同, 主频就可能不同。很多厂商会告诉你相应主频信息。

上面的文中也提到了, 不同机型的CPU, 单核性能是存在差异的。



👍 2



我来也

2020-03-09

学完了今天的课程,捣鼓了下公司阿里云上的服务器资源.
发现按量付费虽好,但长久使用太不划算了!!!
同样的配置, 按量付费一个月的费用是按月付费的近3倍.
也就是说,使用时长超过了10天, 就建议用包月模式了.

发现公司的一个k8s集群,如果转换成包月模式, 几乎可以省去2/3的成本.
这每个月可以省多少钱啊!!!



👍 11



罗辑思维

2020-03-09

老师总结真好, 早几年出课程就好了。

3年前在某云区域代理商干了半年销售。这些云主机型号都不好理解。后来明白一个概念所谓「云计算」的计算, 其实就是指CPU和内存, 所以只要关注CPU和内存比值, 底层硬件信息就容易理解了。



潘政宇

2020-03-09

不同规格，只是cpu内存比例不同吗，底层硬件应该一样吧

作者回复：如文中所说，底层硬件会取决于型号和代别，型号后缀也会有影响



leslie

2020-03-09

文章似乎漏了一点-网络的选择：经典网络应当是共享，专有应当是独享。
核心业务中的核心组件一般会独享：系列和型号确实一直没有注意，更多的关注是在比例上。
什么样的比例用在什么样的场景倒是注意过；
1：1的都有看到过，就看服务器的功能定位，测试服务器，什么东西都有；
1：2的比例确实更多的是偏通用型的综合服务器；
1：4或1：8基本上都是中间件存储/数据系统服务器，以及windows服务器。
今天的课程让我明白。确实云计算在某些方面的选择和实体机确实不一样；CPU参照实体机的思路可能就完全走反了。
谢谢老师的分享：期待后续课程的更新以及学习。

作者回复：关于专有网络和经典网络，请等待网络章节的讲述。

共 2 条评论 >



xpxdx

2020-05-21

老师，文中说的“云上虚拟机的单核性能未必相同”，是指vCPU吗，还是指物理的单核。

作者回复：你好，这里是指vCPU。



戴斌

2020-03-20

突发性机器一直没敢在实际使用，看完文章了解还是有些场景可以用，毕竟便宜





赵子棉

2022-07-20

请问老师，为什么有的内存选项存在3.75GB，15GB，13GB，26GB这样的选项，为什么不是2的幂次？



艾利特-G

2020-03-24

我最近给公司产品上线时用了AWS T3.2xlarge。其实我怀疑他们(开发和别的运维)也没有做过严格的性能测试，不过我也不懂专业测试技术，就按照他们要求创建了。本来开发要求的是4C16G，但是AWS中没有非计算型的4C16G，这也正如老师所说的一样。我本来想创建4C16G的计算型实例的，但是另一个主负责人运维说这个计算型实例贵，还不如换成通用型的T3.2xlarge，比计算型的内存还多一倍呢。他还说业务应用对计算性能要求不高，只是会占用多个CPU。

其实行业里大部分人也都是凭经验拍脑袋做决定，每一门技术都可以深似海，我觉得我作为运维出身的，要抓紧时间把云原生技术栈实操一遍，希望能找到个工作踏踏实实做DevOps.

共 1 条评论 >



简约风

2020-03-22

见过生产最大的机型m5.12xlarge，用于查询变化的场景



我来也

2020-03-10

今天在 <https://jimmysong.io/kubernetes-handbook> 看文章,无意中发现文章底部有本专栏的跳转链接.

哈哈

共 2 条评论 >



一步

2020-03-09

aws的虚拟机通用类型还有t3，a1开头的，这后面的数字也代表第几代吗？如果代表第几代那a1第一代的机器还在用的？

作者回复: 是的, AWS命名中字母后的数字一般就是第几代。A1只是代表这个系列的第一代而已, 没有很老哦。



老姜

2020-03-09

r5.24xlarge大数据计算

共 1 条评论 >



怀朔

2020-03-09

常规服务器 基本上都是2c8g 型号 服务端交互 选型及其类型的时候基本可选系列型号不容乐观。 如阿里云推荐二种比较直接 .1、共享级别 2、企业级。二者差距是共享宿主机器的级别不一样。云厂商更新换代比较快。 老师分析时候比较全 但是用的时候基本上要考虑可用区是否有资源? 跟其他产品是否在同一个可用区内。



Helios

2020-03-09

我们公司的业务是tob的, 最开始没有用云的能力。公司内部搭建hadoop集群, 还有测试交付出去的软件包都是用的物理机 (80c500g) 测试的, 这也是导致公司机器cpu的利用率3%, 内存利用率10%左右的原因。

今天看了这篇文章, 后续要把物理机器搞为不同的配置的虚拟机, 提供给不同的业务场景。

共 1 条评论 >

