

22 | 存储成本：如何推算日志中心的实现成本？

2022-12-19 徐长龙 来自北京



天下无鱼

<https://shikey.com/>

《高并发系统实战课》

[课程介绍 >](#)



讲述：徐长龙

时长 18:38 大小 17.02M



你好，我是徐长龙。

前面我们比较过很多技术，细心的你应该发现了，比较时我们常常会考虑实现成本这一项。这是因为技术选型上的“斤斤计较”，能够帮我们省下真金白银。那么你是否系统思考过，到底怎么计算成本呢？

这节课，我会结合日志中心的例子带你计算成本。

之所以选日志中心，主要有这两方面的考虑：一方面是因为它**重要且通用**，作为系统监控的核心组件，几乎所有系统监控和故障排查都依赖日志中心，大部分的系统都用得上；另一方面日志中心是**成本很高**的项目，计算也比较复杂，如果你跟着我把课程里的例子拿下了，以后用类似思路去计算其他组件也会容易很多。

根据流量推算存储容量及投入成本

在互联网服务中，最大的变数就在用户流量上。相比普通的服务，高并发的系统需要同时服务的在线人数会更多，所以对这类系统做容量设计时，我们就需要根据用户请求量和同时在线人数，来推算系统硬件需要投入多少成本。



很多系统在初期会用云服务实现日志中心，但核心接口流量超过 **10W QPS** 后，很多公司就会考虑自建机房去实现，甚至后期还会持续改进日志中心，自己制作一些个性化的服务。

其实，这些优化和实现本质上都和**成本**息息相关。这么说你可能不太理解，所以我们结合例子，实际算算一个网站的日志中心存储容量和成本要怎么计算。

通常来说，一个高并发网站高峰期核心 API 的 QPS 在 **30W** 左右，我们按每天 **8** 个小时来计算，并且假定每次核心接口请求都会产生 **1KB** 日志，这样的话每天的请求量和每天的日志数据量就可以这样计算：

- 每天请求量 = $3600 \text{ 秒} \times 8 \text{ 小时} \times 300000 \text{ QPS} = 8\,640\,000\,000 \text{ 次请求} / \text{天} = 86 \text{ 亿次请求} / \text{天}$
- 每天日志数据量： $8\,640\,000\,000 \times 1\text{KB} \Rightarrow 8.6\text{TB} / \text{天}$

你可能奇怪，这里为什么要按每天 **8 小时** 计算？这是因为大多数网站的用户访问量都很有规律，有的网站集中在上下班时间和夜晚，有的网站访问量集中在工作时间。结合到一人一天只有 **8 小时** 左右的专注时间，就能推导出一天按 **8 小时** 计算比较合理。

当然这个数值仅供参考，不同业务表现会不一样，你可以根据这个思路，结合自己的网站用户习惯来调整这个数值。

我们回到刚才的话题，根据上面的算式可以直观看到，如果我们的单次请求产生 **1KB** 日志的话，那么每天就有 **8T** 的日志需要做抓取、传输、整理、计算、存储等操作。为了方便追溯问题，我们还需要设定日志保存的周期，这里按保存 **30 天** 计算，那么一个月日志量就是 **258TB** 大小的日志需要存储，计算公式如下：

$$8.6\text{TB} \times 30 \text{ 天} = 258 \text{ TB} / 30 \text{ 天}$$

从容量算硬盘的投入

算完日志量，我们就可以进一步计算购买硬件需要多少钱了。

我要提前说明的是，硬件价格一直是动态变化的，而且不同商家的价格也不一样，所以具体价格会有差异。这里我们把重点放在理解整个计算思路上，学会以后，你就可以结合自己的实际情况做估算了。



目前常见的服务器硬盘（8 TB、7200 转、3.5 寸）的单价是 **2300 元**，8 TB 硬盘的实际可用内存为 **7.3 TB**，结合前面每月的日志量，就能算出需要的硬盘个数。计算公式如下：

$$258 \text{ TB} / 7.3 \text{ TB} = 35.34 \text{ 块}$$

因为硬盘只能是整数，所以需要 **36 块** 硬盘。数量和单价相乘，就得到了购入硬件的金额，即：

$$2300 \text{ 元} \times 36 = \mathbf{82800 \text{ 元}}$$

为了保证数据的安全以及加强查询性能，我们常常会通过分布式存储服务将数据存三份，那么分布式存储方案下，用单盘最少需要 **108 块** 硬盘，那么可以算出我们需要的投入成本是：

$$\mathbf{82800 \times 3 \text{ 个数据副本} = 24.8\text{W 元}}$$

如果要保证数据的可用性，硬盘需要做 **Raid5**。该方式会把几个硬盘组成一组对外服务，其中一部分用来提供完整容量，剩余部分用于校验。不过具体的比例有很多种，为了方便计算，我们选择的比例是这样的：按四个盘一组，且四个硬盘里有三个提供完整容量，另外一个做校验。

Raid5 方式中计算容量的公式如下：

- 单组 **raid5** 容量 $= ((n-1)/n) \times \text{总磁盘容量}$ ，其中 n 为硬盘数

我们把硬盘数代入到公式里，就是：

$$((4-1)/4) \times (7.3\text{T} \times 4) = 21.9 \text{ T} = \text{三块 8T 硬盘容量}$$

这个结果表示一组 **Raid5** 四个硬盘，有三个能提供完整容量，由此不难算出我们需要的容量还要再增加 **1/4**，即：

108 / 3 = 36 块校验盘

最终需要的硬盘数量就是 108 块 + 36 块 Raid5 校验硬盘 = 144 块硬盘，每块硬盘 2300 元。
总成本是：

144 X 2300 元 = 331200 元

为了计算方便，之后我们取整按 **33W** 元来计算。

除了可用性，还得考虑硬盘的寿命。因为硬盘属于经常坏的设备，一般连续工作两年到三年以后，会陆续出现坏块，由于有时出货缓慢断货等原因以及物流问题，平时需要常备 **40 块左右的硬盘**（大部分公司会常备硬盘总数的三分之一）用于故障替换，大致需要的维护成本是
2300 元 X 40 = **92000 元**。

到目前为止。我们至少需要投入的硬件成本，就 T 是一次性硬盘购买费用加上维护费用，即
33 + 9.2 = 42W 元。

根据硬盘推算服务器投入

接下来，我们还需要计算服务器的相关成本。由于服务器有多个规格，不同规格服务器能插的硬盘个数是不同的，情况如下面列表所示：

- 普通 1u 服务器 能插 4 个 3.5 硬盘、SSD 硬盘 2 个
- 普通 2u 服务器 能插 12 个 3.5 硬盘、SSD 硬盘 6 个

上一环节我们计算过了硬盘需求，做 Raid5 的情况下需要 144 块硬盘。这里如果使用 2u 服务器，那么需要的服务器数量就是 12 台（144 块硬盘 / 12 = 12 台）。

我们按一台服务器 3W 元的费用来计算，服务器的硬件投入成本就是 36W 元，计算过程如下：

12 台服务器 X 3W = 36W 元

这里说个题外话，同样数据的副本要分开在多个机柜和交换机分开部署，这么做的目的是提高可用性。

根据服务器托管推算维护费用

好，咱们回到计算成本的主题上。除了购买服务器，我们还得算算维护费用。



把 2u 服务器托管在较好的机房里，每台服务器托管的费用每年大概是 **1W 元**。前面我们算过服务器需要 12 台，那么一年的托管费用就是 **12W 元**。

现在我们来算算第一年的投入是多少，这个投入包括硬盘的投入及维护费用、服务器的硬件费用和托管费用，以及宽带费用。计算公式如下：

第一年投入费用 = 42W（硬盘新购与备用盘）+ 36W（服务器一次性投入）+ 12W（服务器托管费）+ 10W（宽带费用）= 100W 元

而**后续每年维护费用**，包括硬盘替换费用（假设都用完）、服务器的维护费用和宽带费用。计算过程如下：

$9.2W（备用硬盘）+ 12W（一年托管）+ 10W（一年宽带）= 31.2W 元$

根据第一年投入费用和后续每年的维护费用，我们就可以算出**核心服务（30W QPS 的）网站服务运转三年所需要的成本**，计算过程如下：

$31.2W \times 2 年 = 62.4W + 第一年投入 100W = 162.4W 元$

当然，这里的价格并没有考虑大客户购买硬件的折扣、服务容量的冗余以及一些网络设备、适配卡等费用以及人力成本。但即便忽略这些，算完前面这笔账，再想想用 2000 台服务器跑 ELK 的场景，相信你已经体会到，多写一行日志有多么贵了。

服务器采购冗余

接下来，我们再聊聊采购服务器要保留冗余的事儿，这件事儿如果没亲身经历过，你可能很容易忽略。

如果托管的是核心机房，我们就需要关注服务器采购和安装周期。因为很多核心机房常常缺少空余机柜位，所以为了**给业务后几年的增长做准备**，很多公司都是提前多买几台备用。之前有的公司是**按评估出结果的四倍**来准备服务器，不过不同企业增速不一样，冗余比例无法统一。

我个人习惯是根据当前流量增长趋势，评估出的 3 年的服务器预购数量。所以，回想之前我们计算的服务器费用，只是算了系统计算刚好够用的流量，这么做其实是已经很节俭了。**实际你做估算的时候一定要考虑好冗余。**



如何节省存储成本？

一般来说，业务都有成长期，当我们业务处于飞速发展、快速迭代的阶段，推荐前期多投入硬件来支撑业务。当我们的业务形态和市场稳定后，就要开始琢磨如何在保障服务的前提下降低成本的问题。

临时应对流量方案

如果在服务器购买没有留冗余的情况下，服务流量增长了，我们有什么暂时应对的方式呢？

我们可以从节省服务器存储量或者降低日志量这两个思路入手，比如后面这些方式：

- 减少我们保存日志的周期，从保存 30 天改为保存 7 天，可以节省四分之三的空间；
- 非核心业务和核心业务的日志区分开，非核心业务只存 7 天，核心业务则存 30 天；
- 减少日志量，这需要投入人力做分析。可以适当缩减稳定业务的排查日志的输出量；
- 如果服务器多或磁盘少，服务器 CPU 压力不大，数据可以做压缩处理，可以节省一半磁盘；

上面这些临时方案，确实可以解决我们一时的燃眉之急。不过在节约成本的时候，建议不要牺牲业务服务，尤其是核心业务。接下来，我们就来讨论一种特殊情况。

如果业务高峰期的流量激增，远超过 30W QPS，就有更多流量瞬间请求尖峰，或者出现大量故障的情况。这时甚至没有报错服务的日志中心也会被影响，开始出现异常。

高峰期日志会延迟半小时，甚至是一天，最终后果就是系统报警不及时，即便排查问题，也查不到实时故障情况，这会严重影响日志中心的运转。

出现上述情况，是因为日志中心普遍采用共享的多租户方式，隔离性很差。这时候个别系统的日志会疯狂报错，占用所有日志中心的资源。为了规避这种风险，一些核心服务通常会独立使用一套日志服务，和周边业务分离开，保证对核心服务的及时监控。

高并发写的存储冷热分离

为了节省成本，我们还可以从硬件角度下功夫。如果我们的服务周期存在高峰，平时流量并不大，采购太多服务器有些浪费，这时用一些高性能的硬件去扛住高峰期压力，这样更节约成本。

举例来说，单个磁盘的写性能差不多是 200MB/S，做了 Raid5 后，单盘性能会折半，这样的话写性能就是 100MB/S x 一台服务器可用 9 块硬盘 = 900MB/S 的写性能。如果是实时写多读少的日志中心系统，这个磁盘吞吐量勉强够用。

不过。要想让我们的日志中心能够扛住极端的高峰流量压力，常常还需要多做几步。所以这里我们继续推演，**如果实时写流量激增，超过我们的预估，如何快速应对这种情况呢？**

一般来说，应对这种情况我们可以做冷热分离，当写需求激增时，大量的写用 SSD 扛，冷数据存储用普通硬盘。如果一天有 8 TB 新日志，一个副本 4 台服务器，那么每台服务器至少要承担 2 TB/ 天 存储。

一个 1TB 实际容量为 960G、M.2 口的 SSD 硬盘单价是 1800 元，顺序写性能大概能达到 3 ~ 5GB/s（大致数据）。

每台服务器需要买两块 SSD 硬盘，总计 24 个 1 TB SSD（另外需要配适配卡，这里先不算这个成本了）。算下来初期购买 SSD 的投入是 43200 元，计算过程如下：

$1800 \text{ 元} \times 12 \text{ 台服务器} \times 2 \text{ 块 SSD} = 43200 \text{ 元}$

同样地，SSD 也需要定期更换，寿命三年左右，每年维护费是 $1800 \times 8 = 14400 \text{ 元}$

这里我额外补充一个知识，SSD 除了可以提升写性能，还可以提升读性能，一些分布式检索系统可以提供自动冷热迁移功能。

需要多少网卡更合算

通过加 SSD 和冷热数据分离，就能延缓业务高峰日志的写压力。不过当我们的服务器磁盘扛住了流量的时候，还有一个瓶颈会慢慢浮现，那就是网络。

一般来说，我们的内网速度并不会太差，但是有的小的自建机房内网带宽是万兆的交换机，服务器只能使用**千兆**的网卡。



理论上，千兆网卡传输文件速度是 $1000\text{mbps}/8\text{bit} = 125\text{MB/s}$ ，换算单位为 $8\text{mbps} = 1\text{MB/s}$ 。不过，实际上无法达到理论速度，**千兆的网卡实际测试传输速度大概是 100MB/s 左右**，所以当我们做一些比较大的数据文件内网拷贝时，网络带宽往往会被跑满。

更早的时候，为了提高网络吞吐，会采用诸如多网卡接入交换机后，服务器做 **bond** 的方式提高网络吞吐。

后来光纤网卡普及后，现在普遍会使用**万兆**光接口网卡，这样传输性能更高能达到 **1250MB/s**（ $10000\text{mbps}/8\text{bit} = 1250\text{MB/s}$ ），同样实际速度无法达到理论值，实际能跑到 **900MB/s** 左右，即 **7200 mbps**。

再回头来看，之前提到的高峰期日志的数据吞吐量是多少呢？是这样计算的：

$$30\text{W QPS} * 1\text{KB} = 292.96\text{MB/s}$$

刚才说了，千兆网卡速度是 **100MB/s**，这样四台服务器分摊勉强够用。但如果出现多倍的流量高峰还是不够用，所以还是要升级下网络设备，也就是换万兆网卡。

不过万兆网卡要搭配更好的**三层交换机**使用，才能发挥性能，最近几年已经普及这种交换机了，也就是基础建设里就包含了交换机的成本，所以这里不再专门计算它的投入成本。

先前计算硬件成本时，我们说过每组服务器要存三个副本，这样算起来有三块万兆光口网卡就足够了。但是为了稳定，我们不会让网卡跑满来对外服务，最佳的传输速度大概保持在 **300~500 MB/s** 就可以了，其他剩余带宽留给其他服务或应急使用。这里推荐一个限制网络流量的配置——**QoS**，你有兴趣可以课后了解下。

12 台服务器分 **3** 组副本（每个副本存一份全量数据），每组 **4** 台服务器，每台服务器配置 **1** 块万兆网卡，那么每台服务器平时的网络吞吐流量就是：

$$292.96\text{MB/s} \text{（高峰期日志的数据吞吐量）} / 4 \text{ 台服务器} = 73\text{MB/S}$$

可以说用万兆卡只需十分之一，即可满足日常的日志传输需求，如果是千兆网卡则不够。看到这你可能有疑问，千兆网卡速度不是 100MB/s，刚才计算吞吐流量是 73MB/s，为什么说不够呢？



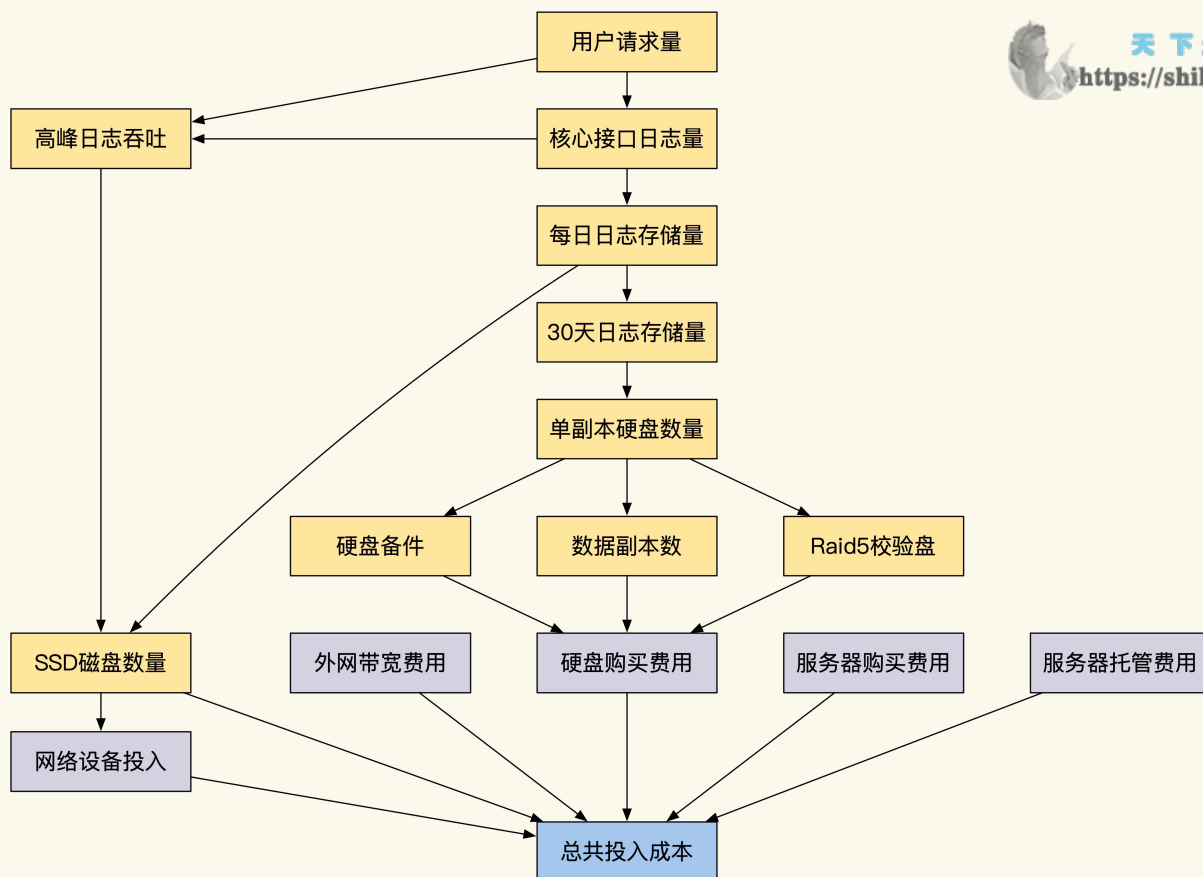
这是因为我们估算容量必须留有弹性，如果用千兆网卡，其实是接近跑满的状态，一旦稍微有点波动就会卡顿，严重影响到系统的稳定性。

另一方面，实际使用的时候，日志中心不光是满足基础的业务使用，承担排查问题的功能，还要用来做数据挖掘分析，否则投入这么大的成本建设日志中心，就有些得不偿失了。

我们通常会利用日志中心的闲置资源，用做限速的大数据挖掘。联系这一点，相信你也就明白了，我们为什么要把日志保存三份。其实目的就是通过多个副本来提高并发计算能力。不过，这节课我们的重点是演示如何计算成本，所以这里就点到为止了，有兴趣的话，你可以课后自行探索。

总结

这节课我们主要讨论了如何通过请求用户量评估出日志量，从而推导计算出需要多少服务器和费用。



推导过程

你可以先自己思考一下，正文里的计算过程还有什么不足。

其实，这个计算只是满足了业务现有的流量。现实中做估算会更加严谨，综合更多因素，比如我们在拿到当前流量的计算结果后，还要考虑后续的增长。这是因为机房的空间有限，如果我们不能提前半年规划出服务器资源情况，之后一旦用户流量增长了，却没有硬件资源，就只能“望洋兴叹”，转而在软件优化方式去硬扛突发 **de** 情况。

当然了，根据流量计算硬盘和服务器的投入，只是成本推算的一种思路。如果是大数据挖掘，我们还需要考虑 CPU、内存、网络的投入以及系统隔离的成本。

不同类型的系统，我们的投入侧重点也是不一样的。比如读多写少的服务要重点“堆”内存和网络；强一致服务更关注系统隔离和拆分；写多读少的系统更加注重存储性能优化；读多写多的系统更加关注系统的调度和系统类型的转变。


尽管技术决策要考虑的因素非常多，我们面临的业务和团队情况也各有不同。但通过这节课，我希望能让你掌握成本推算的思维，尝试结合计算来指导我们的计算决策。当你建议团队自建机房，或者建议选择云服务的时候，如果有一套这样的计算做辅助，相信方案通过的概率也会有所提升。

思考题

1. 建设日志中心，使用云厂商的服务贵还是自己建设的贵？
2. 大数据挖掘服务如何计算成本？

期待你在留言区和我交流互动，也推荐你把这节课分享给更多同事、朋友。我们下节课见！

分享给需要的人，Ta购买本课程，你将得 18 元

 生成海报并分享

 赞 2  提建议

© 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。

[上一篇](#) 21 | 业务缓存：元数据服务如何实现？

[下一篇](#) 23 | 网关编程：如何通过用户网关和缓存降低研发成本？



技术领导力实战笔记 2022

从实操中提升你的领导力

TGO 鲲鹏会

数十位优秀管理者的真知灼见

肖军 / 苏宁金科 CTO
王璞 / DatenLord 联合创始人
郭炜 / 前易观数据 CTO
肖德时 / 前数人云 CTO
林晓峰 / GrowingIO 副总裁
于游 / 马泷医疗集团 CTO
王植萌 / 去哪儿网高级技术总监
胡广寰 / 酷家乐技术 VP
舒超 / 星汉未来 CTO



新版升级：点击「 请朋友读」，20位好友免费读，邀请订阅更有**现金**奖励。

精选留言 (1)

 写留言



陈卧虫 

2022-12-19 来自北京

从前从没想过，原来要花这么多钱

作者回复：技术人员想创业这里也是个槛，刚开始需要用最少的资源做最好的效果

