17-为什么CPU结构也会影响Redis的性能?

你好,我是菲德钧。

很多人都认为Redis和CPU的关系很简单,就是Redis的线程在CPU上运行,CPU快,Redis处理请求的速度 也很快。

这种认知其实是片圈的。CPU的多核架构以及多CPU架构,也会影响到Redis的性能。如果不了解CPU对 Redis的影响,在对Redis的性能进行调优时,就可能会进漏一些调优方法,不是把Redis的性能发挥到极 限。

今天,我们就来学习下目前主流服务器的CPU架构,以及基于CPU多核架构和多CPU架构优化Redis性能的 方法。

丰流的CPU架构

要了解CPU对Redis具体有什么影响,我们得先了解一下CPU架构。

一个CPU处理器中一般有多个全方式。 我们把一个运行核心将为一个物理核、每个物理核都可以运行应用程序、每个物理核能得多机构的一级基件(Level Lache、简称Lache),包括一级指令组存和一级数架接连,以为ABO(2016年)(Level Lache、简称L2 Cache)。

这里提到了一个概念,就是物理核的私有缓存。它其实是指缓存空间只能被当前的这个物理核使用,其他的 物理核无法对这个核的缓存空间进行数据存取。我们来看一下CPU物理核的架构。



因为L1和L2確存是每个物理核私有的,所以,当数据或指令保存在L1、L2維存时,物理核访问它们的延迟 不超过10核的,速度非常快。那么,如果Redis把要退行的指令或存取的数据保存在L1和L2缓存的话,就能 高速地访问S24格令和数据。

但是,这些L1和L2缓存的大小受限于处理器的制造技术,一般只有KB级别,存不下太多的数据。如果L1、

到,是访问11、12维在的研究的近10倍,不可避免给会对性能造成影响。

所以,不到的物理技还会共享一个共同的三级健存(Level 3 cache、脑核为13 cache)。13億有能够使用 的存储资源比较多,所以一般比较大,能达到几MB到几十MB,这就能让应用程序缓存更多的数据。当L1、 12提存中没有数据破存时,可以均同13,尽可能避免访问内存。

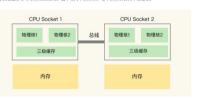
另外,现在主流的CPU处理器中,每个物理核通常都会运行两个超线程,也叫作逻辑核。同一个物理核的逻辑核会共享使用LI、L2條存。

为了方便你理解,我用一张图展示一下物理核和逻辑核、以及一级、二级缓存的关系。



在主放的服务器上,一个CPU处理器会有10到20多个物理核、同时,为了提升服务器的处理能力,服务器 上通常还会有多个CPU处理器(也称为多CPU Socket),每个处理器有自己的物理核(包括L1、L2版 存),13服存,以及连接的内存,同时,不同处理器问题过总线连接。

下图显示的就是多CPU Socket的架构,图中有两个Socket,每个Socket有两个物理核。



段时间, 然后重被调度到Socket 2 上运行。

但是,有个地方需要你注意一下;如果应用程序先在一个Socket上运行,并且把政策保存到了内存,然后 被调度到另一个Socket上运行,此时,应用程序再进行符序访问时,就需要访问之前Socket上连接的内 存,这种访问周于远**端内存访问。和访问Socket直接连接的内存相比,远端内存访问会增加应用程序的延** 况。

在多CPU架构下,一个应用程序访问所在Socket的本地内存和访问远端内存的延迟并不一致,所以,我们 也把这个智构路为主统一内在访问智和(Non-Uniform Memory Access,NUMA智慧)

到这里,我们就知道了主流的CPU多核架构和多CPU架构,我们来简单直结下CPU架构对应用程序运行的影 响。

- L1、L2缓存中的指令和数据的访问速度很快,所以,充分利用L1、L2缓存,可以有效缩短应用程序的执行时间:
- 在NUMA架构下,如果应用程序从一个Socket上调整到另一个Socket上,就可能会出现远端内存访问的情况,这会直接增加应用程序的执行时间。

接下来,我们就先来了解下CPU多核是如何影响Redis性能的。

CPU多核对Redis性能的影响

在一个CPU核上运行时,应用程序需要记录自身使用的软硬件资源信息(例如核指针、CPU核的寄存器值等), 我们把这些信息对**进行时信息**。同时,应用程序访问最频繁的指令和数据还会被缓存到L1、L2模 在 L,以增维对4位行政

但是,在多核CPU的场景下,一旦应用程序需要在一个新的CPU核上进行,那么,运行时信息就需要重新加 截到斯的CPU核上。而且,新的CPU核的L1、L2模存也需要重新加载数据和指令,这会导致程序的运行时 同维加。

说到这儿,我想跟你分享一个我曾经在多核CPU环境下对Redis性能进行调优的案例。希望借助这个案例, 帮你全方位地了解到多核CPU对Redis的性能的影响。

当时,我们的项目需求是要对Redis的99%尾延迟进行优化,要求GET尾延迟小于300微秒,PUT尾延迟小于500微秒。

可能有同学不太清楚99%底延过是啥,我先解释一下。我们把所有请求的处理延迟从小到大排个序,99% 的请求延迟小子的值就是99%尾延迟。此切识,我们有1000个请求,假设按请求延迟从小到大排序后,第 991个请求的延迟实消偿是1ms,而前990个请求的延迟都小于1ms,所以,这里的99%尾延迟就是21ms。

剔开她的时候,我们使用GET/PUT复杂度为O(1)的String类型进行数据存取,同时关闭了RDB和AOF,而 且,Redis实例中没有保存集合类型的其他数据,也就没有bigkey操作,避免了可能导致挺迟增加的许多情 '9-

但是,即使这样,我们在一台有24个CPU核的服务器上运行Redis实例,GET和PUT的99%尾延迟分别是504 微秒和1175微秒,明显大干我们设定的目标。 后来,我们仔细检测了Redis实例运行时的服务器CPU的状态指标值,这才发现,CPU的context switch次数 Https://

context switch是指线程的上下文切换,这里的上下文就是线程的运行时信息。在CPU多核的环境中,一个 纬程先在一个CPU核上运行,之后又切掉到另一个CPU核上运行。这时就会发生context switch。

当context witch发生后。Redis主线程管运行时信息需要被重新加载到另一个CPU核上,而且,此时,另 一个CPU核上的LL 12個存中,并没有Redis实例之前运行对旅客访问的指令和政政。所以,这些指令和政 报客需要重新从13個存,甚至是内存中加载。这个重新加载的过程是需要花式一定针响的。而且,Redis实 例需要等待这个重新加载的过程完成后,才能开始处理排水。所以,这仓会身业一些填弃的处理时间增加。

如果在CPU多核场景下, Redis实的被频繁调度到不同CPU等上距离的话。那么,对Redis实例的请求处理时间移购或更大了。每期度一次,一些课求就会罗斯廷行时信息。指今晚就重新加载过程的影响,这就会导致某些解析的运动用显离于其他请求。分析则这里,我们说知道了附购的例子中99%尼延迟的值始终释不下来的原因。

所以,我们要避免Redis总是在不同CPUN上来回测度执行。于是,我们要试着把Redis实例和CPU标绑定 了,让一个Redis实例固定运行在一个CPUNT上。我们可以使用**taskset带令**把一个程序绑定在一个核上运 行。

比如说,我们执行下面的命令。就把Redis实例绑在了0号核上,其中,"-c"选项用于设置要绑定的核编号。

```
taskset -c 0 ./redis-server
```

绑定以后,我们进行了测试。我们发现,Redis实例的GET和PUT的99%尾延迟一下子就分别降到了260微秒 和482微秒,达到了我们期望的目标。

我们李看一下继续的后的Portis的99%尼亚识。

命令	未绑核运行Redis的99%尾延迟	绑核运行Redis的99%尾延迟
GET	504us	260us
PUT	1175us	482us

可以看到,在CPU多核的环境下,通过绑定Redis实例和CPU核,可以有效降低Redis的尾延迟。当然,绑核 不仅对降低尾延迟有好处,同样也能降低平均延迟、提升香吐率,进而提升Redis性能。

接下来,我们再来看看多CPU架构,也就是NUMA架构,对Redis性能的影响。

CPU的NUMA架构对Redis性能的影响

在实际应用Redis时,我经常看到一种做法,为了提升Redis的网络性能,把操作系统的网络中断处理程序和

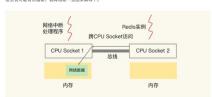
但是,网络中断程序是要和Redis实例进行网络数据交互的,一旦把网络中断程序绑核后,我们就需要注意 Redis实例是排在哪个核上了,这会关系到Redis访问网络数据的效率高低。

我们先来看下Redi实得和网络中断程序的数据交互:网络中断处理程序从网卡配件中设取数据,并把数据 写入到操作系统内核维护的一块内存缓冲区。内核会通过epoll机制能发事件,通知Redis实例,Redis实例 重把物理从内结构的内在维治权维印则自己的内容中间。如下网所示:



那么,在CPU的NUMA架构下,当网络中断处理模型、Redi实例分别和CPU核绑定后,就会有一个潜在的 风险: 如果网络中断处理概序和Redi实例各自所销的CPU核本在同一个CPU Socket上,那么,Redis实例 该取解做界域时,就需要则CPU Socket访问内在。这个讨解企作等的多时间。

这么说可能有点抽象, 我再借助一张图来解释下。



可以看到,愿中的网络中新处理程序被绑在了CPU Socket 1的某个核上,而Redis实例则被绑在了CPU Socket 1上,此时,网络中斯处理程序读取到约网络按照,被保存在CPU Socket 1的本地内容中,当Beelis 实例要访问网络按照时,就需要Socket 2通过总线把内存访问命令发送到 Socket 1上,进行远程访问,时间 开稿补100+

我们曾经做过测试,和访问CPU Socket本地内存相比,跨CPU Socket的内存访问延迟增加了18%,这自然 会导致Redis处理请求的延迟增加。

所以,为了避免Redis跨CPU Socket访问网络数据,我们最好把网络中断程序和Redis实例绑在同一个CPU Socket上,这样一来,Redis实例就可以直接从本地内存读取网络数据了,如下图解示:



不过,要更注意的是,在CPU的NUMA探询下,对CPU接价编号规则,并不是先把一个CPU Socket中的所 有逻辑核编究,再对下一个CPU Socket中的逻辑核编码,而是先给每个CPU Socket中每个物理核的第一 个逻辑核胶次编号,再始每个CPU Socket中的物理核的第二个逻辑核胶次编号。

我给你举个例子。假设有2个CPU Socket,每个Socket上有6个物理核,每个物理核又有2个逻辑核,总共 24个逻辑核。我们可以执行**iscpu命令**,查看到这些核的编号:

可以看到,NUMA node0的CPU核编号是0到5、12到17。其中,0到5是node0上的6个物理核中的第一个逻辑核的编号。12则17是相应物理核中的第二个逻辑核编号。NUMA node1的CPU核编号规则和node0一样。

所以,在绑核时,我们一定要注意,不能想当然地认为第一个Socket上的12个逻辑核的编号就是0到11。否

则,网络中断程序和Redis实例就可能绑在了不同的CPU Socket上。

比如说,如果我们把网络中新程序和Redis实例分别绑到编号为1和7的CPU核上,此时,它们仍然是在2个 CPU Socket上,Redis实例仍然需要跨Socket读取网络数据。

所以,你一定要注意NUMA架构下CPU核的编号方法,这样才不会绑错核。

我们先简单地总结下侧别学习的内容。在CPU多核的场景下,用taskset命专把Bedt实例和一个核绑定,可 以这0%edi实例在不同性上被来回调度执行的开格。避免官高的底证法。在多GPUSWOMA架构下,如果 你对网络中部程序做了维核操作,建议你同时把Redis实例和网络中部程序将在同一个CPU Socket的不同核 上、发程可以增全Bedistrocket的信息在如由服务的整确的指示于44.

不过,"硬币都是有两面的",绑核也存在一定的风险。接下来,我们就来了解下它的潜在风险点和解决方 室。

继核的风险和解决方案

Redis除了主线程以外,还有用于RDB生成和AOF重写的子进程(可以回顾看下<mark>第4讲和_{第5讲})。此外,我们还在第16讲学习了Redis的启台数据。</mark>

当我们把Redis实哪样到一个CPU逻辑核上时,就会导致子进程、后台线程和Redis主线程竞争CPU资源,一 且子进程或后台域特占用CPU的,主线程就会被用塞、导致Redis请求延迟增加。

针对这种情况,我来给你介绍两种解决方案,分别是一个Redis实例对应绑一个物理核和优化Redis源码。

方案一: 一个Redis实例对应绑一个物理核

在给Redis实例绑核时,我们不要把一个实例和一个逻辑核绑定,而要和一个物理核绑定,也就是说,把一个物理核价2个逻辑核都用上。

我们还是以别才的NUMA架构为例,NUMA node0的CPU核编号是0到5、12到17。其中,编号0和12、1和 13、2和14每都是表示一个物理核的2个定栅核。所以,在原核时,我们使用属于两一个物理核的2个逻栅核 进行师核操作。例如,我们执行下面的命令,就把Redis实例绑定到了逻辑核0和12上,而这两个核正好都 属于物理核1.

taskset -c 0,12 ./redis-server

和只將一个逻辑核相比,把Redis演解稱的理核糊定,可以比主核程、予进程、后台核提共享使用2个逻辑 核,可以在一定程度上缓解CPU景海竞争。但是,因为只用了2个逻辑核,它们相互之间的CPU竞争仍然还 合存在、如果你到第一步运心CPU竞争、表面给你介绍一种方案。

方案二: 优化Redis源码

这个方案就是通过修改Redis源码。把子进程和后台线程绑到不同的CPU核上。

如果你对Redis的源码不太熟悉,也没关系,因为这是通过编程实现绑核的一个通用做法。学会了这个方 零,你可以在熟悉了源码之后把它用上,也可以应用在其他需要排核的场景中。

接下来,我先介绍一下通用的做法,然后,再具体说说可以把这个做法对应到Redis的哪部分源码中。

通过编程实现排核时,要用到操作系统提供的1个数据结构cpu_set_t和3个函数CPU_ZERO、CPU_SET和 sched setaffinity,我先来解释下它们。

- cpu set t数据结构; 是一个位图。每一位用来表示服务器上的一个CPU逻辑核
- CPU_ZERO函数:以cpu_set_t结构的位图为输入参数,把位图中所有的位设置为0。
- CPU_SET函数:以CPU逻辑核编号和cpu_set_t位图为参数、把位图中和输入的逻辑核编号对应的位设置为1。
- sched_setaffinity函数: 以进程/线程ID号和cpu_set_t为参数,检查cpu_set_t中哪一位为1,就把输入的ID号所代表的进程/线程排在对应的逻辑核上。

那么,怎么在编程时把这三个函数结合起来实现继核呢?很简单,我们分四步走就行。

- 第一步: 创建一个cpu set t结构的位图变量;
- · 第二步: 使用CPU ZEROIR数。把cpu set t结构的位图所有的位都设置为0;
- · 第三步:根据要绑定的逻辑核编号,使用CPU_SET函数,把cpu_set_t结构的位图相应位设置为1;
- 第四步: 使用sched_setaffinity函数,把程序绑定在cpu_set_t结构位图中为1的逻辑核上。

下面、我就具体介绍下、分别把后台终程、子进程绑到不同的核上的做法。

先说后台线程。为了让你更好地理解编程实现绑核,你可以看下这段示例代码,它实现了为线程绑核的操 作:

```
// MEMARS
- (1985年 1997年 (1985年 1985年 1998年 19
```

对于Redis来说,它是在bio.c文件中的bioProcessBackgroundJobs函数中创建了后台线程。 bioProcessBackgroundJobs函数类似于刚刚的例子中的worker函数,在这个函数中实现缩核四步操作,就 和给线程绑核类似,当我们使用fork创建子进程时,也可以把刚刚说的四步操作实现在fork后的子进程代码 中,示例代码如下:

对于Redis来说、生成RDR和ADF日本重写的子讲程分别是下而两个文件的函数中定现的。

- rdb.c文件: rdbSaveBackground函数;
- aof.c文件: rewriteAppendOnlyFileBackground函数。

这两个函数中都调用了fork创建子进程,所以,我们可以在子进程代码部分加上绑核的四步操作。

使用海际低化方案,我们既可以实现Redis实例继续、避免切换线带来的性能影响。还可以让于进程、后台 线程和主线程不在同一个核上运行,避免了它们之间的CPU資源竞争。相比使用taskset都核来说,这个方 签可以进一步降低绑核的风险。

小结

这节课,我们学习了CPU架构对Redis性能的影响。首先,我们了解了目前主流的多核CPU架构,以及 NUMA架构。

在多核CPU架构下,Redis如果在不同的核上运行,就需要根据地进行上下文切换,这个过程会增加Redis的 执行时间,客户端也会观察到较高的尾延迟了。所以,建议你在Redis运行时,把实例和某个核期定,这 样,能能重复利用帐上的UL)之继存,可以降低响应延迟。

为了並升RedisS向网络性態。提口有时还会把网络电新处理程序和CPU核碳定。在这种情况下,如果服务器 使用的是NUMA架构,Redis实等一旦衰損度到中市新处理程序不在同一个CPU Socket,就更即CPU Socket 访问网络数据,这就会海低Redis的性能。所以,我建议你把Redis实例和网络中新处理程序绑在同一个CPU Socket下的不同核上,这样可以提升Redis的连行核能。

虽然绑核可以帮助Redis降低请求执行时间,但是,除了主线程,Redis还有用于RDB和AOF重写的子进程, 以及4.0版本之后提供的用于惰性删除的后台线程。当Redis实例和一个逻辑核绑定后,这些子进程和后台线 程会和主线程竞争CPU资源,也会对Redis性能造成影响。所以,我给了你两个建议:

- 如果你不想修改Redis代码,可以把按一个Redis实例一个物理核方式进行绑定,这样,Redis的主线程、 之证明和后台体股河以共享体用一个物理核上的两个逻辑核
- 如果你很熟悉Redis的源码,就可以在瞬码中增加IMP核操作,即于速程和后台核复级到不同的核上,这样可以避免对主线程的CPU消溃资金。不过,如果你不熟悉Redis海码。也不用太担心,Redis 6.0世来后,可以支持CPU核级定的配置操作了、我将不知选择中向你介绍Redis 6.009最新特殊。

Redis的低延迟是我们永恒的追求目标,而多核CPU和NUMA架构已经成为了目前服务器的主流配置,所 以、希望你能觉哪個特性化方案、并即它应用到应路中。

每课一问

按照惯例, 我给你提个小问题。

在一台有2个CPU Socket (每个Socket 8个物理核) 的服务器上,我们部署了有8个实例的Redis切片集群(8个实例都为主节点、没有主备关系) 现在有两个方案:

在同一个CPU Socket b 為行8个字例, 并和8个CPU核绑定;

2. 在2个CPU Socket上各运行4个实例,并和相应Socket上的核绑定。

如果不考虑网络数据读取的影响,你会选择哪个方案呢?

欢迎在留言区写字你的思考和答案,如果你觉得有所收获,也欢迎你帮我把今天的内容分享给你的朋友。我 们下转通见。

精选留言:

Kaito 2020-09-16 00:07:35

这篇文章改获很大!对于CPU结构和如何绑线有了进一步了解。其实在NUMA架构下,不光对于CPU的绑 核需要注意,对于内存的使用,也有很多注意点,下面回答误后问题,也会提到NUMA架构下内存方面的 注意事项。

在一台有2个CPU Socket(每个Socket 8个物理核)的服务器上,我们部署了有8个实例的Redis切片集群 (8个实例都为主节点,没有主备关系),采用哪种方案绑核最佳?

我更倾向于的方案是:在两个CPU Socket上各运行4个实例,并和相应Socket上的核绑定。这么做的原因 主要从L3 Cache的命中率、内存利用率、避免使用到Swap这三个方面考虑:

- 1.由于CPU Socket,和2分别有自己的1.5 cache,如果把所有实得都是在两一个CPU Sockett.,相当 于这些实例共用这一个1.3 Cache,另一个CPU Socket的1.3 Cache读贵了。这些实例共用一个1.2 Cache,会等低Cache中的核据解准被接,从同由中事下降,还是只是从为中中核或数据,这些知识可同约2.8 。而为个条例分解定CPU Socket,可以先分使用2个1.3 Cache,提高1.3 Cache的命中事,减少从内存读 对数据的干部。从用语标识见
- 2、如果这些实例都绑定在一个CPU Socket,由于采用NUMA废构的原因。所有实例会优先使用这一个节点的内存。当这个节点内存不足时,再经过总线去申请另一个CPU Socket下的内件。此时也会增加延迟而8个实例分别使用之个CPU Socket,各自在访问内对邻移最处近时间。还好喜此。

3、如果这些实例都现在一个CPU Socket. 还有一个比较大的风险器: 用BISwap的商業再会大大提高, 如果这个CPU SocketV对应的内容不够了,也可能不会走另一个节点申请内存(操作系统可以配置内存 回收馆解SISwap的使用场间: 本节后国现存存 [其管节点申请内存/内存放使用SiSwap的增向管理》,则 作系统可能会把这个节点的一部分内存就能到TSwap上从用释放出内存地进程处用(如果没开启Swap可 会等复重任000 ,因为他的证券更往往来第二,如果SISWap中就收费。 进行电话的性性就会急胜下 降,就还变大,所以8个实务外到现实CPU Socket,既可以充分使用之节点的内存,提高内存使用率, 而且被增值器/Socket) 验检合格性

其实我们可以拿一下,在NUMA架构下,也经常发生某一个节点内存不够,但其他节点内存充足的情况下 ,依旧使用到了Swap,进而导致软件性能急制下降的例子。所以在进伸展面,我们也需要关注NUMA架 将下的内容把情况(多个内存节点使用印能不均衡),并合理能是系统多数《内存图改集版/Swap使用 创造)。尽量未享受中期等wan。(18

etest 2020-09-16 08:49:53

课后问题: 我会选择方案二。首先一个实例不止有一个线程需要运行,所以方案一肯定会有CPU竞争问题 ; 其次切片集群的通信不是通过内存,而是通过网络IO。[1赞]

土豆白菜 2020-09-16 23:10:29
 老师,我也想问下比如azure redis 集香做这些优化

· 洲弋云州 2020-09-16 15 29 04

有两套房子,就不用旅程運吧,优选方案二。老师实验用的X86的CPU吧,对于ARM架构来讲,存在着跨 DIE和跨P的说法,跨P等访问时延会更高,且多个P之间的访问存在着NUMA distances的说法,不同的布 局导致的跨Ps访问时是也不相同。

zhou 2020-09-16 14:54:52

在 NUMA 架构下,比如有两个 CPU Socket: CPU Socket 1和 CPU Socket 2,每个 CPU Socket 都有自己的内存,CPU Socket 1有自己的内存 Mem1,CPU Socket 2 有自己的内存 Mem2。

Redis 实例在 CPU Socket 1 上执行,网络中断处理程序在 CPU Socket 2 上执行,所以 Redis 实例的数据 在内存 Mem1 上,网络中断处理程序的数据在 Mem2上。

因此 Redis 实例读取网络中断处理程序的内存数据(Mem2)时,是需要远端访问的,比直接访问自己的 内存数据(Mem1)要慢。

那时刻 2020-09-16 10:25:05

请问老师,您文中提到我们仔细检测了 Redis 实例运行时的服务器 CPU 的状态指标值,这才发现,CPU 的 context switch 次数比较多。再遇到这样的问题的时候,排查的点有哪些呢?

写点時間 2020-09-16 08:54:20

请问得老师,次曾的例子代码是硬轭的了子进程绑定的CPU编号,这样因为不知道运行时主进程绑定的C PU还是会有一定竞争的风险。那么有没有可以避免这种情况的方案,能够动态根据主进程绑定的情况分 数子进程应该使用的CPU编号的实现呢?

jacky 2020-09-16 08:08:38
 绑核用的都是逻辑核编号吧,那么在云虚机进行相关操作也是一样的了?