拉斯维加斯赌徒神操作,从此SSD性能稳如山!

路向峰 SSDFans Yesterday

点击上方蓝色ssdfans带你进入 万物存储、万物智能、万物互联 闪存2.0新时代



1961年,29岁的年轻人爱德华·索普收到了一个不寻常的邀请,黑道K 先生因为看过他发表的一篇文章《21点的有利策略》,想让他去赌场表 演一下如何战胜庄家。这个邀请不禁让他喜忧参半。喜的是受到认可, 忧的是一旦输了没法向K先生交代;要是赢太多,也会面临赌场的责 难。但是不管怎么说,天使轮的启动资金对于这个年轻人太重要了。

经过慎重考虑,他把验证"商业模式"的地点选择在了雷偌市。索普一战成名,仅仅用了一个周末,就把天使投资额翻了一番。这个消息后来横扫博彩社区,随后,索普写了《战胜庄家》。

索普后来转战华尔街,被人尊称为量化对冲基金之父。量化对冲通过概率统计的方法,实现稳定投资收益和降低投资波动风险的目的。实际上涉及到数量分析和波动抑制的策略也可以用在其他方面。在企业级 SSD 中,垃圾回收和其他后台任务都会受到用户业务类型、强度的变化发生波动。

由于 SSD 的后端写入总带宽固定,垃圾回收效率的波动会造成写入带宽的 抖动。为了更好的理解这一问题,我们尝试从概率的角度入手进行分析,并 采用理论结合仿真的方法给大家呈现了这里的情景。

第一站 拉斯维加斯

大峡谷,胡佛大坝,米德湖等奇特的自然风光勾勒出了这个城市的一面,而令人深陷其中不由贪恋的纸醉金迷般的城市风情则代表了它的另外一面。拉斯维加斯,这座位于沙漠之中的繁华城市,是一座以赌博业为中心的旅游、购物、度假的世界知名度假城市,拥有"世界娱乐之都"和"结婚之都"的美称。

拉加维加斯大道是这座沙漠之城繁华的开始,是不夜城的中心,也是最繁华的地带。街区两旁是无比奢华的酒店、热闹非凡的赌场。还有各种埃及金字塔、埃菲尔铁塔、自由女神像这样的世界奇迹的缩影让你目不暇接。走在熙熙攘攘的街道,你可以看热情奔放的金发女郎,呼啸而过的豪车。



拉斯维加斯的主要经济支柱是博彩业,在大道街上有高级的旅馆赌场和秀场。不少游客本着猎奇的态度,走进这个虚幻的世界。而赌场则利用对于人性地把握,让赌客在享受到视听盛宴服务的同时,心甘情愿的把自己的钱投入到一个个赌局当中。而赌场赢钱的秘密,则在于利用人性的特点设计不公平的赌局。即便是看似"公平"的赌博,实际上也会最终让赌客输得精光。在这个奥秘的背后,起作用的是概率学的客观规律。

概率学在长久的发展中为我们提供了许多有趣的定理,其中就有那么一条赌徒输光定理。根据赌徒输光定理,在"公平"的赌博中,任一个拥有有限赌本的赌徒,只要长期赌下去,必然有一天会输光。在一次赌博中,任意一个赌徒都有可能会赢,谁输谁赢是偶然的。但只要一直赌下

去,输光或者庄家破产跑路是必然的。然而确实有人曾经利用数学知识打败了庄家,并取得了"赌博教父"的称号。

他就是文章开头故事中《战胜庄家》的作者——爱德华·索普。索普通过他的21点制胜理论,一战成名。一时间,坊间巷里流传着这样的故事:有个人"奇袭"了内华达雷偌市的所有赌场,并成功地从"21点"桌上赢了数万美元。那些曾经遭遇横扫的赌场,纷纷将其拉入黑名单,并合谋更改游戏规则……索普还间接推动了期权定价模型的发展,最终导致著名的Black Scholes期权定价模型的诞生。

接下来我们的故事将从布朗运动开始,借助概率论数学知识,提出赌徒筹码的期望模型。模型中反应的规律也许对于大家理财有一些价值,不过我们还是会把落脚点放在SSD如何保证恒稳的写入带宽。我们知道,在SSD完全顺序写入的时候,SSD的写放大可以低到1,而顺序写入的带宽会很高。而在随机写入的时候,由于垃圾回收的存在,写放大会增加,带宽则会降低。

实际上,存在一些情况,使得SSD会面临比纯随机写更加困难的局面。 而对于SSD的设计者来说,要保证各种情况下的带宽和带宽的稳定,就 需要和用户施加的IO Pattern这个庄家进行一次次博弈,并通过猜测这 个庄家的底牌来达到优化服务的目的。

第二站 爱因斯坦的论文与布朗运动

时间退回至 1905 年,年仅 26 岁的爱因斯坦还在瑞士伯尔尼专利局任三级技术员,这一年他似乎突然天才迸发,连续发表了 4 篇引发科学革命的论文。其中有关光电效应,质能等效性和狭义相对论的3 篇论文为世人熟知,还有一篇是有关布朗运动的。

在这篇论文中,爱因斯坦指出:在空间中,如果一个质点在时刻 0位于原点,每过单位时间该质点从当前位置沿随机方向移动单位距离到一个新的位置,那么这个质点的运动服从布朗运动,并且在时刻n它到原点的距离的平均值满足 $\sqrt{\mathbb{E}(d^2)} = \sqrt{n}$ 。

我们把这个结论例化到一维的情况。假设一个赌徒参加一个公平的赌局,他每次有一半的概率赢 1 元,一半的概率输 1 元。这样 n 次以后,他的筹码数量变化的期望值为 $\sqrt{E(d^2)} = \sqrt{n}$,这说明随着时间增加,赌徒手里的筹码数量相比初始筹码数量的差值将会越来越大。最终或者赌徒赢到足够多的钱不再继续赌下去,或者是输光了黯然离场。然而根据赌徒输光定理,很有可能出现的是后一场景。

为了说明这个结论,假定赌徒手里初始有r元,且赌徒一直不停赌博直到输光为止,我们尝试求解输光的次数小于 n 的概率 p(r,n)。由于每次最多输 1 元,因此至少需要 r 次,才能出现输光的情况。当的时候,每次或者赢 1 元,或者输 1 元,因此只有当 n-r 为偶数的情况,才会出现恰好输光的情况。因此

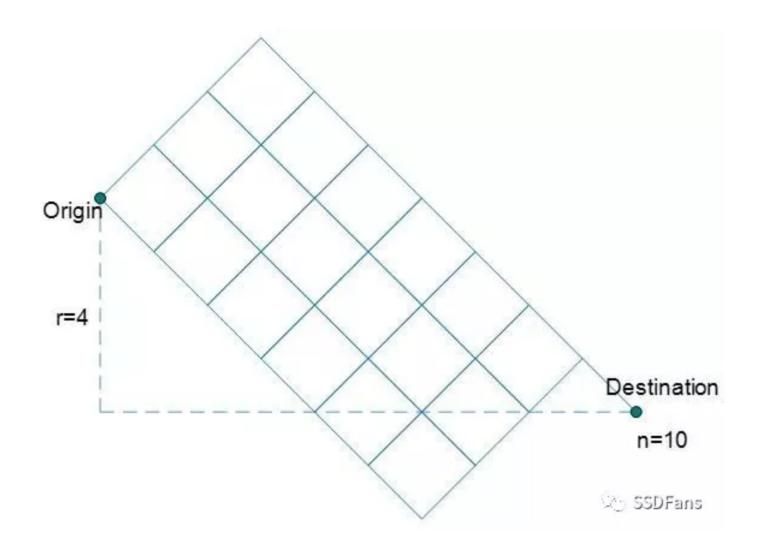
$$\mathbf{p}(r,n) = 0, \qquad n < r$$

$$\mathbf{p}(r,n+1) = \mathbf{p}(r,n), \qquad n = r + 2m(m \text{ is an integer})$$

采用古典概率公式,为了求解 n-r 为非负偶数时的 p(r,n),设 n-r=2m,我们把所有的可能情况分成三种:

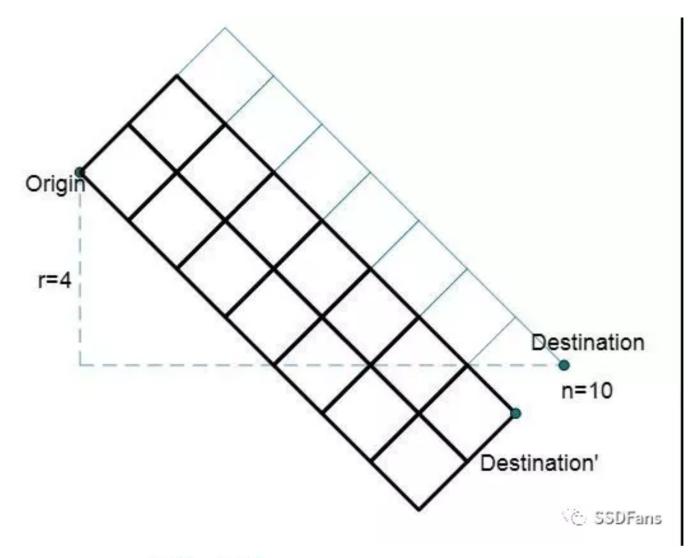
- E1:截至时刻n,赌徒手中金额一直大于零;
- (2) E2:恰好在时刻n,赌徒输光;
- 3 E3:在时刻n之前赌徒已经输光,并且扩散到时刻n的各种可能性。

容易知道 $\mathbf{p}(r,n) = (\|E2\| + \|E3\|)/2^n$ 。下面分别计算 $\|E2\|$ 和 $\|E3\|$ 。为了计算E2事件的次数 $\|E2\|$,考虑下面 r=4,n=10 的情况,从原点 Origin 到终点 Destination 有不同的路径(图中以实线表示)。用户每一步从网格中的一个顶点向右上或者右下到下一个定点。由于 E2 表示在第 n 步,赌徒恰好输光,所以在往右上或者右下进行行走的过程中不能与 y=0 这条虚线相交。最后一步恰好从坐标点 (9,1) 进入到 (10,0) 满足恰好输光。



为了求解 $\|E2\|$,把它转化成从原点 Origin 到终点Destination,并通过坐标点 (n-1, 1) 的所有的路径数 c_1 减去与y=0这条线相交的路径数 c_2 ,即 $\|E2\|=c_{e2}(n,m)=c_1-c_2$ 。前者容易计算为 $c_1=C_{n-1}^{m-1}$ 。为了计算后者,我们考虑所有与 y=0 这条线相交,并且至少有一处相交后往右上的方向行走。

当m>0的时候,利用镜像原理,把这处往右上行走的步骤改为往右下行走,那么最终结束的位置将会是坐标点 Destination' (9,-1) 。容易证明采用上述做法进行镜像的路径与那些与 y=0 这条线相交的路径——对应,因此 $c_2 = \frac{C_m-1}{2}$,如下图所示。



而一旦知道了 $\|E^2\|$, $\|E^3\|$ 可以把各种在n时刻前已经输光的情况合并计算得到:

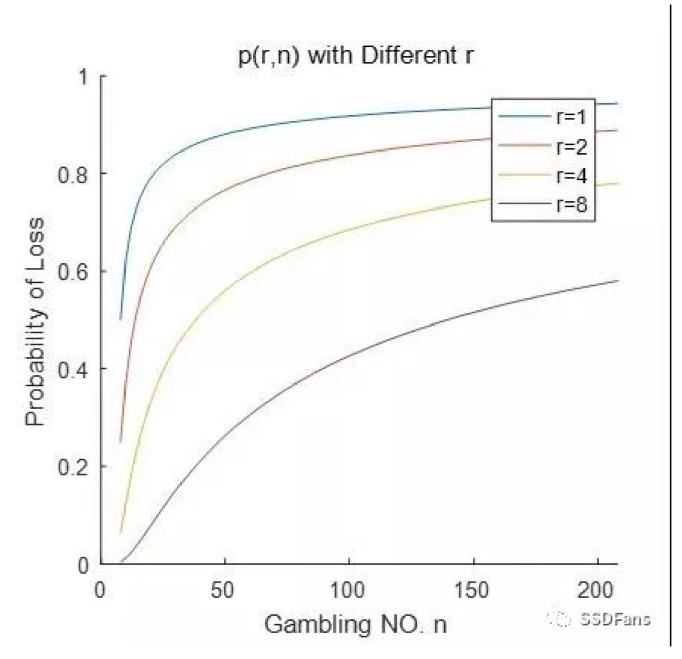
$$||E3|| = c_{e3}(n,m) = \sum_{i=0}^{m-1} c_{e2}(r+2i,i) \ 2^{2(m-i)}$$

综合起来(定义 $C_{n-1}^{-1}=0$),

$$\mathrm{p}(r,n) = \frac{(\|E2\| + \|E3\|)}{2^n} = 2^{-r} \sum_{i=0}^{(n-r)/2} \left(C_{r+2i-1}^i - C_{r+2i-1}^{i-1} \right) 2^{-2i}$$

我们把不同 r 对应的 p(r,n) 放到同一个图中进行比较。它形象地揭示了赌徒输光定理的含义,显然"公平"赌博并不公平。从结果上看,随着 n 的增加,赌徒输光的概率会逐渐增加并趋近于 1 ,并且r越

小,这种趋势越明显。这说明在公平赌博的情况下,拥有筹码最少的 赌徒会更容易破产。



对于一个理性的玩家赌徒, 必须试图在每次赌局具有高于一半的胜算,这种赌博才是有收益的。我们修改一下上述问题,每次赌博的胜率是 α ,失败概率是 β =1- α 。在 α >0.5 的情况下,每次收益的期望值是正数,因此可以预期这个赌博游戏会源源不断的盈利。事实的真相是这样的吗?

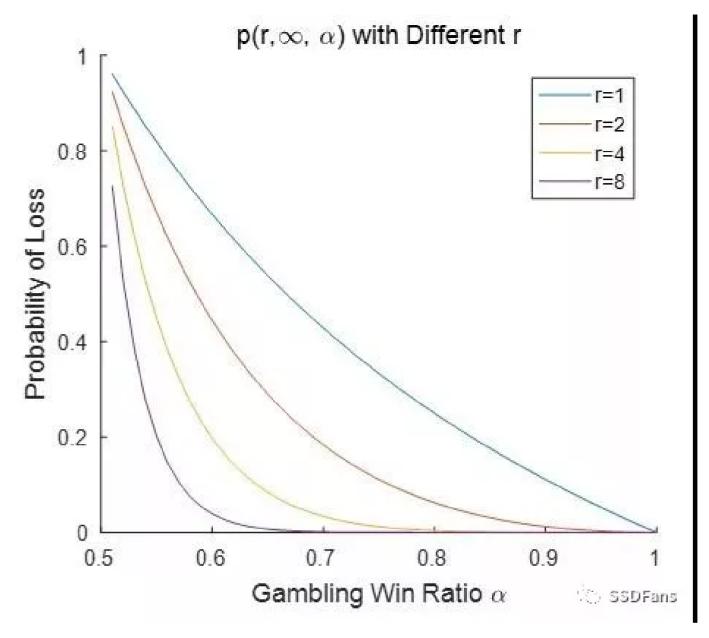
由于每次赌博都具有随机性,虽然每次的期望值是正数,但是还是有一定可能性会出现输光离场的情况。不过,随着赌徒逐渐赢的钱逐渐增加,这种概率会越来越小。为了定量说明这种情况,考虑这样一个不断赌博的过程,设随机变量 \mathbf{T} ,代表赌徒输光离场的时间, \mathbf{T} ,代表赌徒 一旦赢了。元之后就决定离场的时间。设 $\mathbf{p}(r,s,\alpha) = P(\mathbf{T}_r = min(\mathbf{T}_r,\mathbf{T}_s))$ 代表赌徒因为输光而离场的概率。考虑每次赌博胜率是 α ,失败的概率是 β ,那么可知

$$p(r,s,\alpha) = \alpha p(r+1,s-1,\alpha) + \beta p(r-1,s+1,\alpha)$$

这个式子是二阶线性差分方程,带入初值 $p(0,s,\alpha)=1$, $p(r,0,\alpha)=0$,可以得到

$$p(r, s, \alpha) = \frac{1 - (\alpha/\beta)^s}{1 - (\alpha/\beta)^{s+r}}$$

为了讨论赌徒输光的概率,令 $s \to \infty$,得到 $p(r, \infty, \alpha) = (\alpha/\beta)^{-r}$ 。我们也把不同的r对应 $p(r, \infty, \alpha)$ 画出来。从结果上看,如果希望输光的概率比较小,那么需要每次的赢面足够大或者是手里的筹码足够多。对于如果 α 接近0.5,为了保持持续盈利,必须手里筹码足够多才可行。



如果初始资金是r,但是每次赌注是2元,那么输光的概率又是多少呢?容易知道,这等价于初始资金是r/2和每次赌注是1元的情况。在这种情况下,虽然每次的期望赢钱数量翻倍了,但是输光的概率确是按照平方根方式增加了。在r不高的情况下输光的概率会大大增加。

我们的讨论印证了索普的理念:永远不要多下注,如果你过度下注,即使你有优势,你也可能输光。值得一提的是我们还能够从讨论中得到一些有趣的推论,包括马太效应以及富人事业更加容易成功。我们在选择

投资性商品或的时候需要平衡风险和收益,在玩德州扑克游戏的时候则需要根据自己和对手的筹码来选择策略。那么SSD的设计和研发与这些讨论有什么关系呢?

第三站 SSD中的垃圾回收

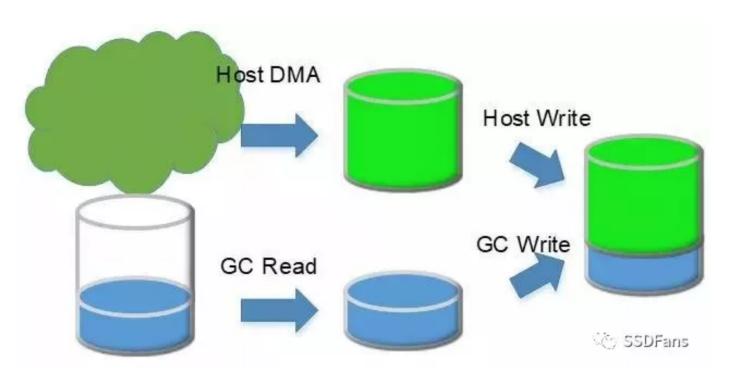
我们已经从赌徒输光定理谈到了爱因斯坦的论文,有了这些知识的准备,我们现在来看看SSD中的垃圾回收。

根据《深入浅出SSD 固态存储核心技术、原理与实践》这本书的介绍,由于Flash需要先擦再写的特性,因此需要在SSD的OP(Over Provision,即SSD中为了方便垃圾回收会多预留一些空闲块)资源即将耗尽的时候,选取那些有较少数据的数据块(称为Victim Block),并且把这个数据块上的数据读出来并写入到新的位置。这个过程称为垃圾回收。

通常 Victim Block 的选择采用贪心算法,即选择有最少数据的数据块。也会适当考虑其他因素,例如在 "Design Tradeoffs for SSD Performance" 这篇论文中提到的基于时间的淘汰算法,适当考虑了冷数据并优化磨损均衡。

在下面的图中,选择出来的Victim Block中有效数据比例为 v ,那 么 SSD 会把有效数据逐步从 Victim Block 中读出来并且写入到新 的位置。同时会节约出来 1-v 对应的 Block 的空间,这个空间可以 用来写入用户数据。在这个过程中,设 q 代表 GC 写入数量与总写入数量的比值,如果保持 GC 写入数量占比 q=v ,那么使得GC回

收完成一个 Block 的时候,刚刚好把新使用的 Block 相同大小的空间释放出来,从而维持了OP资源不变,实现了垃圾回收的持续进行。若 q=v,则对应的写放大为 1/(1-v)。在SSD的实现中,可以实现一个灵活调整用户写入数据的强度和GC的强度的流控机制,来实现 q=v 的目标。



但是,上述的过程实际上是一个过于理想化的过程。这是因为:

- v 代表了 Victim Block 中有效数据比例,它是一个估计的值,可能会与实际情况有误差;
- a于 Victim Block 的选择是选择有最少数据的数据块和其他磨 损均衡算法综合考虑的结果,因此,不同的 Victim Block 对应 的v并不相同。如果始终保持 q=v ,那么会导致垃圾回收的速度 发生波动,进而导致用户写入带宽的波动;

3

由于用户业务工况也可能会发生变化,造成上述持续平稳的回收过程会被破坏。

为了解决上述问题,需要在系统中预留一部分空闲的 Block 资源,当不同的 Victim Block 的有效数据发生波动的时候,这些资源可以用来补偿和平滑这部分波动。这样使得垃圾回收的强度不会随着v而剧烈抖动,以及 SSD 表现出来的性能尽量平稳,避免OP 资源耗尽从而 SSD 无法进行垃圾回收。

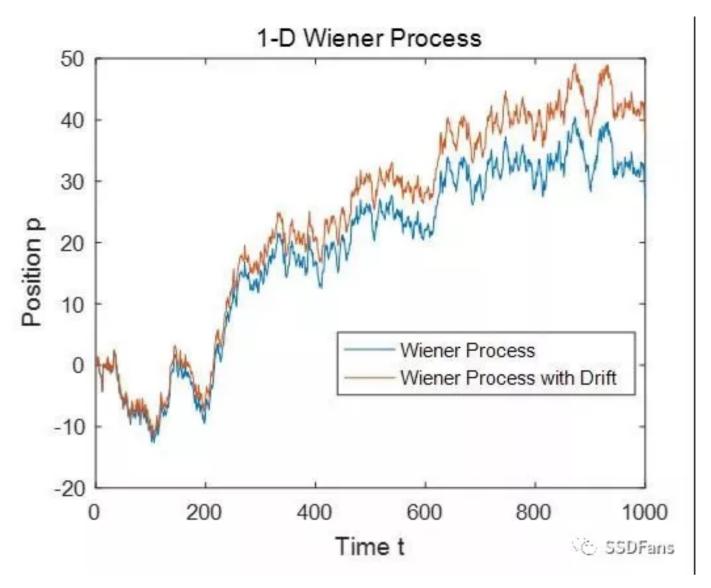
设这部分资源的数量为r,那么问题来了,到底r是多大合适?如果r较小,在r一旦耗尽,正常的垃圾回收就无法进行,导致严重后果;反之,过大的r则会浪费宝贵的OP,引起写放大的增加。此外,在一定的r的前提下,到底如何选择q,才能在波动性和安全性之间找到平衡点?如果采用保守的策略,可以选择大一些的q,可以避免耗尽资源但同时会导致更大的波动;反之更加激进策略则会导致资源耗尽的概率增加,从而造成严重的问题。实际上,上述关于赌徒策略的分析可以帮助我们更好地理解这个问题。

"赌徒"SSD与"庄家"I/O请求的博弈

最初SSD手上有r个Block的筹码,一旦它把这些筹码全都用光了,那么SSD再也不能找到一个空闲的Block执行垃圾回收,从而系统完全不能工作。而它的对手庄家则是用户发送的IO请求,庄家可以给SSD发送不同的命令来干扰垃圾回收过程,从而导致垃圾回收效率下降和造成写请求带宽的抖动。SSD的底牌则是在SSD内部所有含有效数据的Block的状态。

一旦庄家知道了这些SSD的底牌,庄家则可以有针对性地向特定位置 发送写请求,使得这些Block中含有有效数据的分布呈现更加复杂的 态势,这会导致SSD选择Victim Block更加困难和波动更大。庄家甚至可以"恶意地"干扰磨损均衡过程,造成SSD内部的Block寿命分布呈现扩大化趋势。

所幸的是,庄家并不完全知道SSD内部Block中有效数据分布的状态。对于SSD来说,IO 业务并不是完全对立的关系。因此,我们可以假定这个"庄家"是善意的。这种善意最好的表示就是假设虽然 Victim Block的有效数据量v存在波动,呈现一定的随机性,但是同时v仍然满足一定的分布,可以用一个概率分布函数来表示。进一步地,可以假设 v 服从 $N(\mu,\sigma^2)$ 的正态分布。为了尽量保持垃圾回收的速度不变,我们保持 q 在一段时间里是常量。为了说明r随时间变化的规律,我们引入连续时间的布朗运动,即维纳过程。



数学中,维纳过程是一种连续时间随机过程。如果t>s,那么维纳过程 要求在时刻t的位置 W_t 和s时刻的位置 W_s 满足 $W_t-W_s^{N(0,\sigma^2(t-s))}$ 。在上述SSD进行垃圾回收的模型中,如果设垃圾回收的相对比例 $q=\mu$,那么r随Victim Block的回收过程近似满足维纳过程即 $r=r_0+W_t$ 。由于 W_t 在任意时刻均值为零,所以r的均值也不会变化,但是r的方差会随时间逐渐变大。

可以类比赌徒输光定理,采用上述的方案,在 $q=\mu$ 的情况下,如果时间足够长,r 耗尽的概率趋近于 1,也就是说如果垃圾回收的速度等于平均的有效数据的比例的情况下,只要时间足够长,总是会遇到空闲块资源r耗尽的情况。因此必须使 $q>\mu$,才能使得系统在长时间足够安全。在这种情况下 $\Gamma=(q-\mu)t+r_0+W_t$ 是一个上鞅。引用

"Adventures in stochastic processes" 中有关上鞅停时的结论,这种情况下长时运行导致r耗尽的概率为

$$p(q, \mu, r_0, \sigma) = e^{-\frac{2(q-\mu)r_0}{\sigma^2}}$$

根据上述公式,在v的分布 $N(\mu, \sigma^2)$ 不变的情况下,空闲块资源 r_0 多,那么系统安全性越高。为了保证一定的概率 p 下对应的 q ,可以把上述公式求解q的表达式为

$$q = \mu - \sigma^2 \ln(p)/(2r_0)$$

这个式子解释了在一定的r资源耗尽概率下,资源数r,v的分布参数 μ 和 σ^2 与当前预期的垃圾回收比例之间的函数关系。由于 p<1,从而 $q>\mu$,说明在进行垃圾回收的时候由于 Victim Block 中有效数据量的抖动,必须保证垃圾回收速度要略微快于有效资源数的均值 μ ,才能实现比较安全的垃圾回收。在进行决策的时候,是根据当前可用的资源数r而不是 r_o 来进行控制的。

与此同时,由于 q>μ,采用上面的方式有可能会导致潜在资源数量 r发散的情况。这会导致预留的空闲资源侵占 OP 的空间。为了解决这个问题,可以设置一个资源数量的上限 r_m。并且同样利用上述方法构造一个下鞅限制系统以大概率不会出现资源数超出上限的情况。综合上下界之后的公式为

$$q = \mu + \frac{\sigma^2 \ln(p)}{2} \left(\frac{1}{r_m - r} - \frac{1}{r} \right)$$

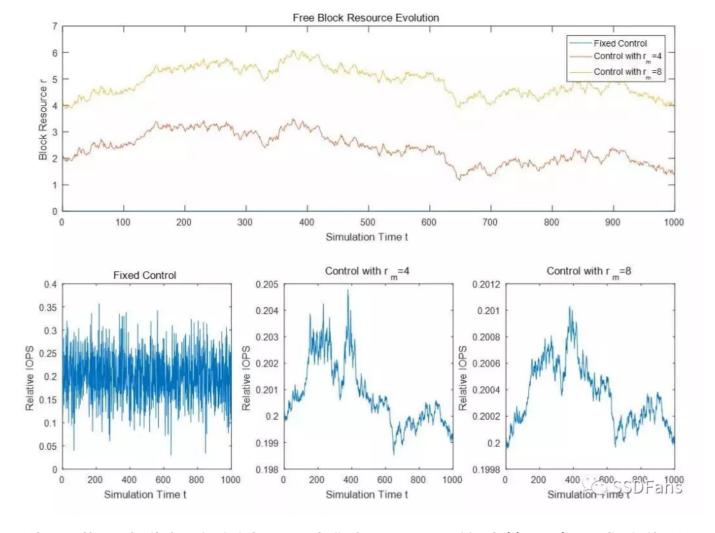
最后一站 让实践给这场讨论画上句号

我们比较和仿真了三种不同的情况。

- 固定采用q=v的方案。这种情况并不需要空闲的Block资源。所以在这种情况下可以认为r恒等于零;
- ② 设定 r_m=4,这种情况下SSD最多可以使用8个空闲Block作为缓冲;
- ③ 设定 r = 8 , 可以预期这种情况下的稳定性更好。

针对三种情况,设置相同的 p , μ 和 σ^2 并仿真资源数r随时间的变化和相对IOPS(定义成用户写入带宽占后端写入总带宽的比例,等于1-q)随时间的变化规律。仿真结果如下图所示。从结果可以看出,后两种方法有效实现了系统的资源数r控制到了 0 到 r_m 这 r 个区间范围内。

同时可以看到,采用固定 q=v 的控制方法会使得 IOPS 抖动很厉害。而采用资源池缓冲的方法,会让 IOPS 显得平滑。另外 r_m越大, IOPS 的波动越小。因此,采用上述方法控制后,SSD 是有信心赢得这场赌局的,即采用一定量的空闲 Block 资源池来提升系统的 IOPS 写入的稳定性。



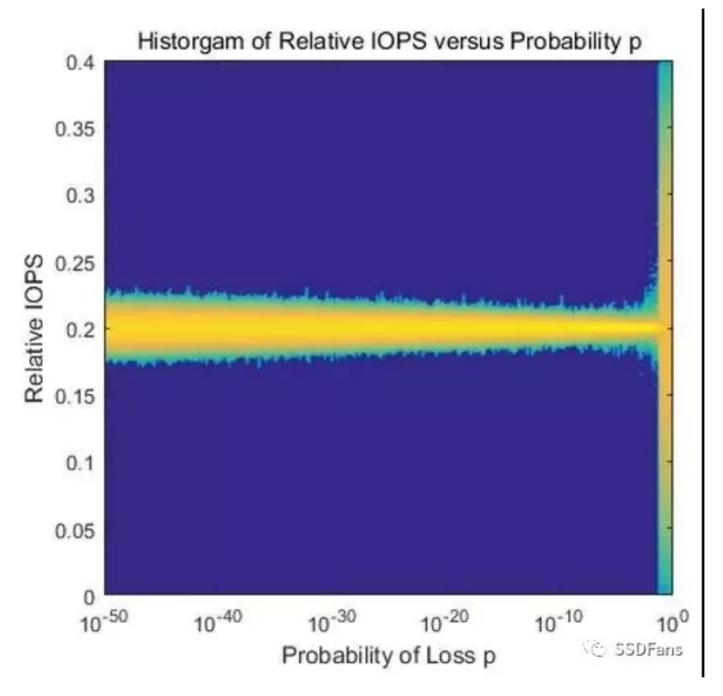
为了进一步分析这个过程,我们把r和q的计算公式写成迭代公式:

$$q_{n} = \mu + \frac{\sigma^{2} \ln(p)}{2} \left(\frac{1}{r_{m} - r_{n}} - \frac{1}{r_{n}} \right)$$
$$r_{n} = r_{n-1} + 1 - \frac{v_{n}}{q_{n-1}}$$

上面的迭代公式表明, r_n , q_n 仅与 r_{n-1} , q_{n-1} 有关,这符合马尔可夫过程。因此, r_n , q_n 构成了连续状态的马尔可夫链。我们可以通过MCMC方法来估算它的稳态分布,进而研究不同的参数对于系统的稳定性的影响。当 σ^2 足够小和 r_m 足够大情况下,进行近似后,可以得到q的方差满足下面的关系。

$$var(q) = \frac{-16\sigma^4 ln(p)}{\mu r_m^2}$$

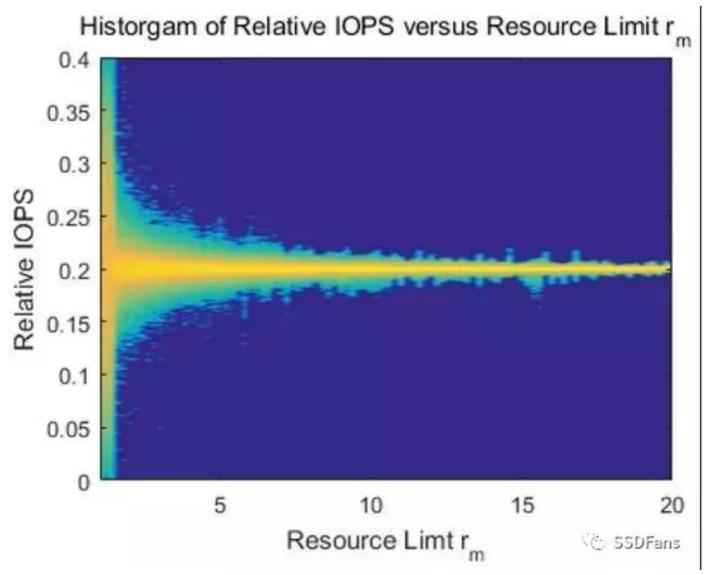
我们把不同的参数 p 下,相对的 IOPS 的分布概率密度直方图画出来。从图中可以看出来,当p比较小的时候,由于ln(p)这一项计算出来的绝对值比较大,因此系统控制的相对比较保守。如果出现r偏移中心值较大的情况,会非常灵敏地调整IOPS以稳定r,因此相对的IOPS分布的比较宽(在图中反映的是对应的概率下黄色部分的高度较高)。



随着p逐渐增加,IOPS的分布会逐渐收窄,这说明我们适当放宽约束,可以使得相对的IOPS更加稳定,这与上述理论计算结果吻合。不过当p增加到10⁻³的时候,可以看到IOPS的分布会突然变宽。这是因为随着控制策略的逐渐松弛,会有越来越大的概率出现r接近于 0或者是 r_m的情况。一旦出现这种情况,就会导致控制策略的强烈反应。这会使得 IOPS 不得不尽量考虑保证 r 不会超限,从而牺牲

IOPS的稳定性。因此p的选择决定了短期波动和长期波动的平衡。p 选的比较大有助于保证短时间内的抖动较小,但是却可能会影响到r 的变化趋势,造成长期波动的增加。因此要根据实际情况来进行选 择。

此外,固定p,改变 r_m 也可以理解空闲Block的数量对于IOPS抖动性的影响。这个所反应的规律与直觉是一致的,即随着 r_m 的增加,IOPS的抖动性呈现反比关系下降的规律,这与上述公式反应的q的方差与不是反比例关系的规律一致。因此在条件允许的情况下,可以适当增加一些空闲块资源用于平滑IOPS的波动。



在上面的分析中,通过对于 Victim Block 的建模,通过估算Victim Block短时分布规律,我们从上得出了一个可信的垃圾回收稳定性方法,这个方法会对于SSD整体的写入带宽的平稳起到重要作用。实际情况中,对于 μ 和 σ^2 需要有一个估计的过程,需要平衡估计的准确度和估计的跟踪速度之间的关系。

把这个方法推广开来,采用有效的方法来预测未来长时的垃圾回收效率,可以不仅仅保证短期的稳定性,也可以保证在 IO 业务发生剧烈变化的时候SSD的垃圾回收尽量符合预期和平稳化,限于篇幅关系这里不再展开论述。

随着SSD的广泛应用,业界对于IO的稳定性和服务质量的关注日益加强。而存储的池化则为深度挖掘数据中的特性并进行优化提供了广泛的可能性。紧密结合应用方式并充分利用SSD的能力的设计将会极大改变现有的存储架构并成为存储创新的源动力。

转载自Memblaze微信公众号

高端微信群介绍

创业投资群

AI、IOT、芯片创始人、投资人、分析 师、券商

闪存群	覆盖5000多位全球华人闪存、存储芯片精英	
存储群	全闪存、软件定义存储SDS、超融合等企业级存储	
AI芯片群	讨论AI芯片和GPU、FPGA、CPU异构计算	
物联网群	物 联网、5G技术与产业讨论	
ASIC群	芯片和FPGA硬件技术讨论群	
存储芯片群	DRAM、NAND、3D XPoint等各类存储介质和主控讨论	
CIO群	企业信息化讨论	
渠道群	存储产品报价、行情、渠道、供应链	

想加入这些群,长按或扫描下面二维码加nanoarchplus为微信好友,介绍你的姓名-单位-职务,注明群名,拉你进群。

Read more