华中科技大学

研究生课程考试答题本

考生姓名	
考生学号	
系、年级	武汉光电国家实验室 2017级
类别	非定向
考试科目	计算机系统分析与性能评价
考试日期	2017年12月5日

Model and Analysis of RAID Storage System

Shan Kai

***Abstract*—RAID,Redundant Array of Independent Disks is a data storage virtualization technology that combines multiple physical disk drive components into one or more logical units for the purposes of data redundancy, performance improvement, or both. But how to quantify the evaluation of RAID performance is a problem. In this paper, a concise mathematic tool is adopted, Queueing Model, to establish a system model of RAID. By the general RAID model, the performance of RAID storage system can be measured, which give a theoretical basis for further studying on the optimal performance of RAID storage system.**

***Keywords*-RAID;Queueing model;M/G/1; Performance; Reliability**

I. Introduction

RAID, Redundant Array of Inexpensive (or Independent) Disks is a data storage virtualization technology that combines multiple physical disk drive components into one or more logical units for the purposes of data redundancy, performance improvement, or both.

Storage system performance, especially access speed, is extremely important for computers and networks.The improvement of system performance depends on the correct analysis method and careful calculation and design. Only through systematic analysis to identify the key factors that affect performance, and thus take measures to overcome the weak links in the system to achieve the balance of all links in order to achieve the desired speed requirements. RAID is widely used for its superior features, but the speed of storage systems is far behind the speed of CPU development. We need to analyze the performance of RAID, so find ways to improve RAID performance.

The contributions of this paper is: This paper establishes the system model of RAID0,RAID1 and RAID5 based on M/G/1 queueing theory and build model for FC-RAID in SAN.

This paper is organized as follows. Section Ⅱ presents RAID ,FC-RAID and some statistics for disk drives. Section Ⅲ establishes a model of RAID0,RAID1,RAID5 and FC-RAID using M/G/1 model. In Section Ⅳ, we present some analysis result, followed by the conclusion in Section Ⅴ.

II. Preliminary

A. RAID and RAID level

RAID mainly thanks to the block and cross-access technology. Disk array performance differences because of the adoption of different data cross granularity and redundant information placement and calculation methods, so by the data organization is divided into different levels. All levels are in line with the following three things in common, only RAID0 level no parity information, not subject to article 3).

1) RAID is a collection of drives (either HDDs or SSDs or a mixture of both), but is considered a logical drive device under the operating system.

2) data block, and distributed in a group of disk drives

3) redundant drive used to store the verification information, in the event of a drive failure to ensure data recovery.

RAID0 Its settings are for performance comparison with other levels. At the same time because of its high data rate, but for some of the backup can be reissued, requiring data transmission rate of high occasions. It has the best performance due to the calculation of non-redundant information, but the data is less reliable.

RAID1 Each of its data disks has a mirrored disk. When you write to a disk drive, the data is written to its mirrored disk at the same time; while the read operation, only the disk drive with the shortest wait time and the shortest seek time is read. When a disk fails, it can read data from its mirror disk. High reliability, high I / O rates, double storage capacity, and high cost.

RAID5 Disk arrays whose parity information's distribution is rotated use block interleave.

RAID usually has seven levels, here we analyze three.

B. FC- RAID

Fiber Channel Disk Array (FC-RAID) storage systems are composed of embedded software and hardware. FC-RAID hardware architecture mainly include the following components: processor, memory, FC card, SCSI card,SCSI disk. Fiber card work in the target mode, acts as master-slave channel card, used to receive the host to send I / O requests. The processor processes the SCSI commands according to the RAID level. The SCSI commands are forwarded to the SCSI card. The SCSI card works in the initiator mode and acts as a serial controller. The SCSI commands are executed in parallel between the strings so that multiple disks can do I/O operation simultaneously, return data and status to the host finally. In the storage area network based on Fiber Channel disk array technology, users access the storage server through the IP network, and the server forwards the user request to the Fiber Channel Disk Array for data access through the Fiber Switch. When multiple hosts access Fiber Channel Disk Arrays at the same time, their data access patterns characterize like I/O requests in the network environment.

I/O process in the storage area network is as follows: The storage server sends I/O requests to the disk array through Fiber Channel. The I/O requests are queued at the end of the array in master-slave channel card. When a read request is made, search for the corresponding data, the read hit, directly from the prefetch cache to read data back to the host, if not hit, then read prefetch command will be generated, read data from the disk and write to cache, and the data into return to the queue waiting for Fiber Channel writing back to the host. For the write command, using write and wear strategy, the data is written to the disk that the write completed, while returning to the host to complete the state.

C. statistical average of disk drive access time

Disk drive access time is part of the subsystem I / O response time, which contains three parts of time: seek time, wait for sector arrival time, data transfer time, if T_{seek} , T_{rot} ,and T_d , respectively, the driver's Access time T_{access} is the sum of these three, that is

$$\mathrm{T}_{access} = T_{seek} + T_{rot} + T_d \tag{1}$$

where $\mathrm{T}_\mathrm{d} = B / r_d$ (B is the amount of data transferred, r_d is the disk drive data transfer rate).

Seek Time T_{seek} The probabilistic distribution of seek time is related to the application environment of the host computer and the realization technology of the disk seek actuator, which can generally be regarded as uniform distribution, exponential distribution or polynomial distribution. Here we adopt exponential distribution whose probability density function is:

$$\mathrm{f}_s(x) = \frac{1}{s} e^{-\frac{x}{s}} \qquad x \geq 0 \tag{2}$$

s is a constant related to disk characteristics.The average and second order moments of T_{seek} are:

$$\mathrm{E}[T_{seek}] = s \tag{3}$$

$$\mathrm{E}[T_{seek}{}^2] = 2s^2 \tag{4}$$

Calculation process see appendix.

Rotation Delay Time T_{rot} Because the disc speed is constant, the sectors are evenly arranged on the track, and the seek time is independent of the rotation angle of the disc. Therefore, the T_rot obeys the uniform distribution, and in the worst case, equal to the disc rotation time, the best The situation is 0, the probability density function is $\mathrm{f}_r(x) = \frac{1}{N}$,where N is the num of sectors on a track. Therefore, the statistical mean value $\mathrm{E}[T_{rot}]$ and the second moment $\mathrm{E}[T_{rot}{}^2]$ of the waiting sector time are obtained as $\mathrm{E}[T_{rot}] = \frac{1}{2}\mathrm{Nt}$ and $\mathrm{E}[T_{rot}{}^2] = \frac{1}{3}(\mathrm{Nt})^2$. Where, t for a sector turn time, $\mathrm{N_t}$ for a turn of time.

III. Model Analysis

A. The performance analysis of RAID0

RAID0 I / O response time can be obtained from the disk drive access time and control system overhead. Suppose the number of disk drives in the system is N, as shown in Figure 1.

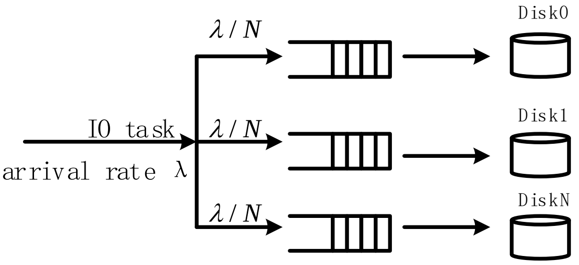

Figure 1 RAID0 Drive Service Queuing Model

If the I / O request arrives according to Poisson distribution, the request arrival rate is a, and after the command is decomposed, a sub-command queue is formed for each driver and the probability of reaching each driver is equal and equal probability distribution, and then each drive command arrives with a probability of 1 / N, each drive requests a reachable rate $\lambda_{\mathrm{i}} = \lambda / N$ 。Each drive in the system has its own data path, and commands can be executed in parallel. For simplicity, assume that read and write operations have the same access time. So you can find out the average I / O service time for each drive in RAID0 is：

$$\mathrm{E}[T] = E[\xi] + E[\varsigma] + B / r_d + B / r_c \qquad (5)$$

where $E[\xi]$, $E[\varsigma]$ are Seek time and waiting sector time expectations respectively，B / r_d ，B / r_c are transfer time of a batch of data B with the drive transfer rate r_d and the SCSI channel's data rate $r_{c..}$ The degree of dispersion of T is expressed in terms of its second moment：

$$\begin{aligned}\mathrm{E}[T^2] &= E[\xi^2] + E[\varsigma^2] + 2E[\xi]E[\varsigma] \\ &+2a(E[\xi] + E[\varsigma]) + a^2\end{aligned} \qquad (6)$$

where $\mathrm{a} = B / r_d + B / r_c$

According to the P-K formula of M / G / 1 queuing model, the average queuing waiting time of each driver can be obtained as

$$\mathrm{T}_{queue} = \frac{\lambda_i E[T^2]}{2[1 - \lambda_i E[T]]} \qquad (7)$$

The average response time should be the average of the I / O service times the disk drive transfers data through the SCSI adapter plus the average queuing latency for each drive, so the average drive-to-array response time, Tdisk, is

$$\mathrm{T}_{disk} = E[T] + T_{queue} = E[T] + \frac{\lambda_i E[T^2]}{2[1 - \lambda_i E[T]]} \qquad (8)$$

Since multiple drivers work in parallel, the average response time of the array $\overline{\mathrm{T}\varpi}$ is, at best, equal to the average response time Tdisk of the drive-to-array,so

$$\overline{\mathrm{T}\varpi} = \mathrm{T}_{disk} = E[T] + \frac{\lambda_i E[T^2]}{2[1 - \lambda_i E[T]]} \qquad (9)$$

If the disk drive's request arrival rate is large enough for the disk drive to work at full capacity, that is, the controller's CPU is fully on standby for a certain period of time, the disk drive utilization $\eta_{\mathrm{disk}} = \lambda_i E[T] = 1$，the maximum throughput of the drive can be calculated as $\lambda_{\mathrm{i\,max}} = 1 / E[T]$.

Therefore, the maximum rate of arrival of the array is

$$\lambda_{\max} = N \square [\frac{1}{E[T]}] \qquad (10)$$

In the case of the maximum throughput, if Bmax is the data transfer rate at this time, then

$$\mathrm{B}_{\max} = \lambda_{\max} \square \mathrm{B} E[L] = \frac{N \square B E[L]}{E[T]} (Byte / s) \qquad (11)$$

where, L is the size of the IO request block, the block size is evenly distributed between [1, W], and the average block size is $\mathrm{E}[L] = (1 + W) / 2$. Degree of dispersion (second moment) $\mathrm{E}[L^2] = (W^2 + W + 1) / 3$.

B. The performance analysis of RAID1

RAID1 is a typical mirror disk structure, its service queuing model shown in Figure 2:

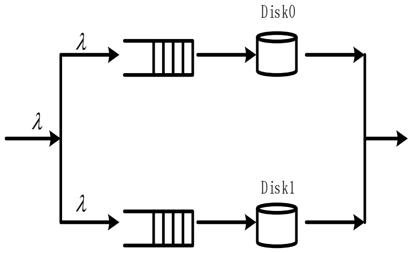

Figure 2 mirror system queuing model

Similarly, suppose the arrival of I / O requests is Poisson distribution, and the request arrival rate is λ. Both drives read and write the same data, and their arrival rates are the same, equal to the total I / O request arrival rate. If the length of the data to be read and written is L, the I / O service time expressed as its expected value is:

$$\mathrm{E}[T]=E[\xi]+E[\varsigma]+\mathrm{L}/r_d+L/r_c \tag{12}$$

The degree of dispersion of expectations expressed as the second-order moment is：

$$\begin{aligned}\mathrm{E}[T^2]&=E[\xi^2]+E[\varsigma^2]+d^2E[L^2]+2E[\xi]E[\varsigma]\\&+2dE[L](E[\xi]+E[\varsigma])\end{aligned} \tag{13}$$

where $\mathrm{a}=1/r_d+1/r_c$.

According to the M / G / 1 queuing model, the average queuing latency T_{queue} for each drive can be calculated as:

$$\mathrm{T}_{queue}=\frac{\lambda_i E[T^2]}{2[1-\lambda_i E[T]]} \tag{14}$$

The average response time $\overline{\mathrm{T}\omega}$ is:

$$\overline{\mathrm{T}\omega}=\mathrm{E[T]}+\mathrm{T}_{\mathrm{queue}}=E[T]+\frac{\lambda_i E[T^2]}{2[1-\lambda_i E[T]]} \tag{15}$$

When disk utilization $\eta_{\mathrm{disk}}=\lambda E[T]=1$，$\lambda_{\max}=1/E[T]$ 。With the number of bytes said，

$$\mathrm{B}_{\max}=\lambda_{\max}\square E[L]=\frac{E[L]}{E[T]} \tag{16}$$

C. RAID5 chunk data request I / O response time analysis

In RAID5, if the total number of drives is N, the number of one verification group is equal to (N-1) blocks. In this way, the data length L of each IO is less than or equal to the data amount of one verification group in terms of the number of blocks and is identified as a large block, $\mathrm{L}\in[1,N-1]$. When less than a verification group of data, after the grouping, disk storage space debris. If equal to a verification group of data, the block, just make use of all the block space. Discussed separately as follows:

1) Fragment read and write service time

When the IO request is a read request, the read service time Tra is: $\mathrm{T}_{ra}=E(\xi)+E[\varsigma]+a$;When the IO request is a write request, the write service time $\mathrm{T_{wa}}$ is: $\mathrm{T}_{wa}=[E(\xi)+E[\varsigma]+a]+[E[\varsigma]+a+d]$, where $\mathrm{d}=2LB\tau$,(L is the number of request blocks, B is the block size, t is the exclusive OR calculation time)

2) Full block read and write service time

When full block, the verification calculation time is $[(N-1)-1]B\tau$, writing service time is: $\mathrm{T_{wb}}=[E(\xi)+E[\varsigma]+a]+(n-2)B\tau$,reading service time is $\mathrm{T}_{rb}=E(\xi)+E[\varsigma]+a$.Suppose the probability of two types of IO service requests are Pa, Pb, and Pa + Pb = 1, then the service time of the driver T is:

$$\mathrm{T}=\mathrm{P_r}(P_aT_{ra}+P_bT_{rb})+P_w(P_aT_{wa}+P_bT_{wb}) \tag{17}$$

Their expected and second moments are E[T] and E [T^2], respectively.

Again, according to the M/G.1 queuing model, $T_{queue}=\frac{\lambda E[T^2]}{2[1-\lambda E[T]]}$, $\overline{T}\omega$=E[T]+T$_{queue}$, $\lambda_{max}=1/E[T]$,and $B_{max}=\lambda_{max}\square E[L]B=\frac{E[L]}{E[T]}B$.

In the above discussion and calculation, the reference probability of Pa, Pb, Pr, Pw, etc., but do not know what IO distribution should be distributed function, according to the literature that the IO load obeys the geometric distribution.

D. The performance analasis of FC-RAID

From the above we can see that the disk array I/O response process can actually be divided into two parts: prefetch scheduling queuing process and Fiber Channel transmission queuing process. For the read request, the input process of prefetch queue is exactly the process of request entry for the Fiber Channel transmit queue, and for the write request, the output of Fiber Channel transmit queue is just the request entry process for the prefetch queue. Therefore, the system's service queuing model can be seen as two queuing systems in series. According to the queuing theory, if the arriving process is Poisson distribution in the tandem queuing system, and the service time of each service station is subject to the independent negative exponential distribution, then each service station can be considered separately. The tandem queuing system can be reduced to a separate queuing system. So we can consider prefetch scheduling service queue and Fiber Channel transmission queue separately. The figure of the model is shown in Figure 3.

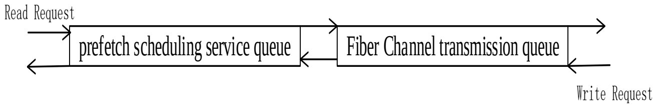
Read Request
prefetch scheduling service queue
Fiber Channel transmission queue
Write Request

Figure 3 FC-RAID service model

1) Prefetch scheduling service time

Prefetch scheduling queuing model is shown in Figure 4.

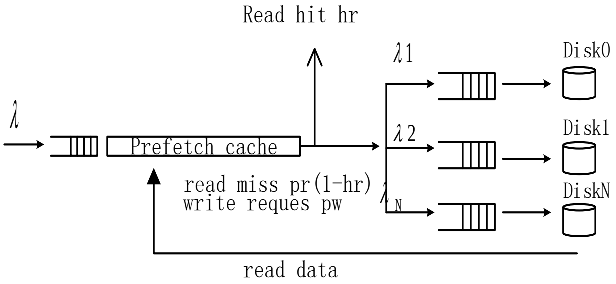

Figure 4 Prefetch scheduling queuing model diagram

Assuming that the I / O request is Poisson distribution, the request arrival rate is λ, where the read request probability is pr, the write probability is pw = 1-pr, the read-hit probability is hr, and the number of disk array disks is N. Each of the N disk drives has an independent data transfer channel and can achieve full parallel operation.

Disk service time, disk seek time, rotation delay and their expectations and second-order moments .etc are the same as above.

Assume that when a read request misses a prefetch cache, I / O operations have the same probability on all disk drives, that is, the load is balanced, and the I / O request rates on each disk drive in the array are equal.

$$\lambda_i=\frac{[p_r(1-h_r)+p_w]\lambda}{N} \quad \text{i=1,2,3}\cdots\text{N} \qquad (18)$$

According to the M / G / 1 queuing formula, the average response time of the I / O request in the disk drive can be obtained as:

$$\mathrm{T}_{disk} = E[T] + \frac{\lambda_i E[T^2]}{2[1-\lambda_i E[T]]} \tag{19}$$

So for the prefetch scheduling service queue, the I / O average response time:

$$\mathrm{T}_{prefetch} = \frac{[p_r(1-h_r)+p_w]\lambda}{N} \cdot T_{disk} \tag{20}$$

2) Fiber Channel transmission time

For Fiber Channel transmission service queues, assuming that the data transmission time is a negative exponential distribution and the average service time is $\mathrm{T_{fc}}$, the average speed of services is μ. According to the M/M/1 queuing formula , the average queueing time of I/O requests can be obtained: $\mathrm{T}_{transfer} = \dfrac{T_{fc}}{1-\lambda T_{fc}}$.

Then the total response time of I/O average Fiber Channel disk array is:

$$\mathrm{T}_{\text{FC-RAID}} = T_{prefetch} + T_{transfer} \tag{21}$$

IV. The Result Analysis

A. Analysis Result of RAID0

According to equation (3) ,(4),(5)and(6), with the same I / O request arrival rate, as the number of disk drives increases, λ_{i} decreases, so average queuing latency decreases, so that the average IO response time of the RAID0 disk array decreases and the system throughput increases. According to equation (6), increasing the block size for I / O requests E[L] increases the maximum data transfer rate for RAID0.

B. Analysis Result of RAID1

The two drives in the mirroring architecture have the same status. The I / O commands do not need to be decomposed. The two drives have the same load size. That is, the data volume processed by each drive is L, so the data traffic per second is equal. As L increases, the I / O response time also increases proportionately.

C. Analysis Result of RAID5

All requests are full block writes with the highest throughput of IOs and the longest response time of all “small write” case.

D. Analysis Result of FC-RAID

The higher the cache hit rate, the shorter the average response time. When the load is all random read, the cache hit rate is the lowest, the response time is longest.

V. Conclusion

RAID storage system is a parallel system, with the characteristics of scalability, transparency and heterogeneity. Queuing theory is a way to analyze RAID performance. With the development of knowledge base, machine learning, intelligent problem solving, parallel database, scientific computing and other fields, RAID will be put forward higher requirements. In this paper, by modeling and analysis several levels of RAID and FC-RAID, obtained some ways to enhance RAID performance.

VI. Appendix

A. The expectation of T_{seek} and the derivation of second moment

$$
\begin{aligned}
\mathrm{E}[T_{seek}] &= \int_0^{+\infty} x\frac{1}{s}e^{-\frac{x}{s}}dx \\
&= \frac{1}{s}\int_0^{+\infty} xe^{-\frac{x}{s}}dx \\
&= \frac{1}{s}[x(-s)e^{-\frac{x}{s}} - \int_0^{+\infty} -se^{-\frac{x}{s}}dx] \\
&= \frac{1}{s}[x(-s)e^{-\frac{x}{s}} + s\int_0^{+\infty} e^{-\frac{x}{s}}dx] \\
&= \frac{1}{s}[x(-s)e^{-\frac{x}{s}} + s(-s)e^{-\frac{x}{s}}] \\
&= \frac{1}{s}[(x+s)(-s)e^{-\frac{x}{s}}]\Big|_0^{+\infty} \\
&= \frac{1}{s}s^2 = s
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[T_{seek}{}^2] &= \int_0^{+\infty} x^2\frac{1}{s}e^{-\frac{x}{s}}dx \\
&= \frac{1}{s}\int_0^{+\infty} x^2e^{-\frac{x}{s}}dx \\
&= \frac{1}{s}[x^2(-s)e^{-\frac{x}{s}} - \int_0^{+\infty} 2x(-s)e^{-\frac{x}{s}}dx] \\
&= \frac{1}{s}[x^2(-s)e^{-\frac{x}{s}} + 2s\int_0^{+\infty} xe^{-\frac{x}{s}}dx] \\
&= \frac{1}{s}[x^2(-s)e^{-\frac{x}{s}} + 2s[x(-s)e^{-\frac{x}{s}} - \int_0^{+\infty}(-s)e^{-\frac{x}{s}}dx]] \\
&= \frac{1}{s}[x^2(-s)e^{-\frac{x}{s}} + 2s[x(-s)e^{-\frac{x}{s}} + s(-s)e^{-\frac{x}{s}}]]\Big|_0^{+\infty} \\
&= \frac{1}{s}[x^2(-s)e^{-\frac{x}{s}} + 2s[(x+s)(-s)e^{-\frac{x}{s}}]]\Big|_0^{+\infty} \\
&= \frac{1}{s}[x^2+2s(x+s)](-s)e^{-\frac{x}{s}}\Big|_0^{+\infty} = 2s^2
\end{aligned}
$$

References

[1] Thomasian A, Han C, Fu G, et al. A performance evaluation tool for RAID disk arrays[C]//Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings. First International Conference on the. IEEE, 2004: 8-17.

[2] Lee E K, Katz R H. An analytic performance model of disk arrays[C]//ACM SIGMETRICS Performance Evaluation Review. ACM, 1993, 21(1): 98-109.

[3] 张江陵，冯丹. 海量信息存储[J]. 北京：科学出版社, 2003, 17.

[4] 周大水，马绍汉. RAID1 的实现策略及性能研究[J]. 计算机研究与发展，1997，34(2): 137-142

[5] Luo X, Li D. The research of mechanical access time of synchronous disk array[C]//Computer Science and Software Engineering, 2008 International Conference on. IEEE, 2008, 4: 198-201.