Study on Performance of IP-SWAN Based on Distributed NS-RAID

Baojiang Cui [1], Jun Liu [2,3], Gang Wang[2], Jing Liu [2]

1. Information Security Centre, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2.Dept. of Computer Science, Nankai University, Tianjin 300071, China; 3. Dept. of Information, Tianjin University of Finance and Economics, Tianjin 300222, China

Email: cui_bj@sina.com.cn

Abstract

Previous work on performance of network storage system has been mostly qualitative. This paper proposes a quantitative analytical method based on closed queueing networks in order to analyze the performance bounds of IP-SWAN (Storage Wide Area Network) based on distributed Network Software RAID (NS-RAID). Experimental results show that testing results are within the bounds predicted by the performance analysis model and the bounds reflect the dynamic trend of the actual testing performance. Furthermore, the bottleneck and the potential bottleneck that affect the IP-SWAN performance can be derived from the performance analysis model. Meanwhile the model provides us the criteria for distinguishing the key performance factors and non-key performance factors.

1. Introduction

Since the distance between host and storage devices connected by LAN or SAN is limited, it is hard to satisfy the requirements for distributed storages across the Wide Area Network (WAN). Furthermore, with the increasing development of applications, such as disaster recovery [1], contents distribution and data grid [2], there is a critical need for data storage within a global scope.

IP-SWAN (Storage Wide Area Network) is a kind of storage network that can implement data access at block level globally by IP storage technique. By IP storage protocols, IP-SWAN can get access to network storage devices within the scope of WAN. IP-SWAN implementations are mainly divided into two types. One is the extension of SAN based on Fibre Channel (FC) within the scope of WAN. Two remote FC-SANs are connected by IP WAN using FCIP bridge or mFCP / iFCP gateway. The other is the storage network in which host gets access to storage devices connected by WAN exclusively through IP network storage protocol (for example iSCSI [3], HyperSCSI [4], ENBD, etc.).

Although many prior works have been done on the performance of distributed network storage systems in the literature, most of them were qualitative [3][5][6]. The quantitative analysis model is still limited [7]. So we focus on quantitative analytical method for IP-SWAN based on the closed queueing networks.

2. IP-SWAN architecture

The IP-SWAN architecture based on distributed NS-RAID is shown in Figure 1. The storage devices distributed in WAN are mapped into virtual disks at local centre server. These virtual disks are organized various levels RAID storage space by software RAID driver at centre server. Thus distributed storage resources across the WAN form a virtual space with a single I/O space, available at local centre server.

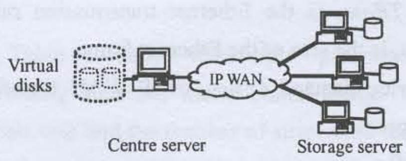


Figure 1. IP-SWAN architecture

The basic data flow of IP-SWAN based on distributed NS-RAID is as follows. The storage devices at storage server side are mapped into the virtual disks at centre server by ENBD through IP WAN. The I/O requests from applications at centre server to local virtual RAID storages are passed to remote storage server via IP WAN by ENBD client. When the ENBD server at storage server receives the packets, it resolves them into original data and I/O commands. Then I/O commands do the particular read or write operation on storage devices through the device file system or device driver. Finally the correspondent acknowledgements are fed back to the centre server.

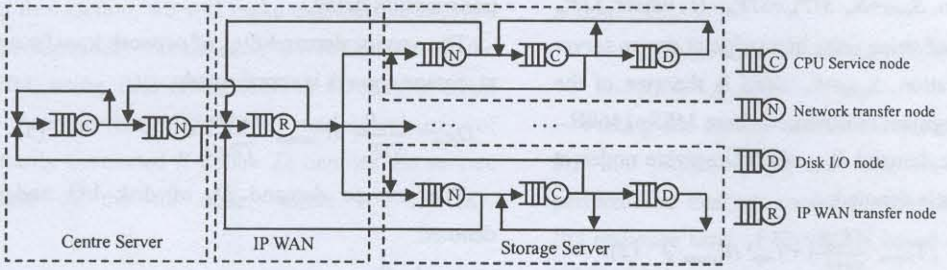


Figure 2. Closed queueing networks model for IP-SWAN

3. CQNM for IP-SWAN

3.1 Closed queueing networks model

A Closed Queueing Networks Model (CQNM) for IP-SWAN based on distributed NS-RAID is established at the basis of Section 2 (see Figure 2). The centre server at left side is connected to storage server at right side by IP WAN. All the main components of IP-SWAN are abstracted service nodes in the queueing networks, including CPU service nodes, network transfer nodes, IP WAN transfer nodes and disk I/O nodes. The CPU service nodes are used for processing local applications and data. Network transfer nodes are used for sending / receiving data through NIC. IP WAN transfer nodes take charge of transferring data through IP WAN. Disk I/O nodes are used to do read / write operations. We assume that each node in Figure 2 is an M/M/1 queue.

3.2 Service demands

For the *i*th node, its service demand is defined as the product of the visit ratio and the mean service time of the *i*th node per job [8].

The service demand D_{Cm} of CPU service nodes at centre server includes the time required by service nodes reading / writing virtual disks, the time spent by ENBD client processing data and the time cost by TCP/IP processing overhead when cache misses. Thus we derive:

$$D_{Cm}= T_{mpro}\cdot \frac{S_m}{STU_m} +(1-P_m)\cdot T_{mp}\cdot IP_{num} \qquad (1)$$

In Equation (1), S_m is the bytes processed by a job; T_{mpro} is the mean process time of a single data stripe unit at centre server (including the time of memory copy and read / write operations); STU_m is the size of stripe unit at centre server; P_m is cache hit probability for read operation; $1-P_m$ represents probability of a job accessing disks at storage servers for read operation. Suppose that cache hit probability is zero for write operation because the file is far more than system cache size. T_{mp} is the sum of the time spent by ENBD client processing data in an IP packet and the time cost by TCP/IP processing overhead. IP_{num} is the number of IP packets into which a job is fragmented, $IP_{num}=\lceil S_{mR5}/MSS\rceil$, where S_{mR5} is the number of bytes processed by centre server for a RAID 5 job. For write operation, $S_{mR5}=S_m\cdot STP_m/(STP_m-1)$, where STP_m is the number of stripe units in a stripe at centre server. For read operation, $S_{mR5}=S_m$. MSS is the size of the largest TCP segment in Ethernet, here MSS=1460B.

The service demand D_{Csn} of CPU service nodes at storage servers is denoted:

$$D_{Csn}=\frac{1-P_m}{STP_m}\cdot\left(T_{snpro}\cdot\frac{S_{sn}}{STU_m}+T_{snp}\cdot IP_{snnum}\right)\quad(2)$$

In Equation (2), T_{snpro} is the mean time spent by storage server processing a stripe unit with the size of STU_m; S_{sn} is the number of bytes processed by storage server for a RAID 5 job from centre server, $S_{sn}=S_{mR5}/STP_m$, T_{snp} is the time spent by storage server processing data in an IP packet. IP_{snnum} is the number of IP packets processed by storage server for a RAID job from centre server, denoted by $IP_{snnum}=IP_{num}/STP_m$.

The service demand D_{Nm} of network transfer nodes at centre server is described:

$$D_{Nm}=(1-P_m)\cdot\frac{IP_{num}\cdot Frames_e}{TRate_e}\quad(3)$$

Where, $TRate_e$ is the Ethernet transmission rate and $Frames_e$ is the size of the Ethernet frame.

The service demand D_{IPWAN} of IP WAN transfer nodes is expressed:

$$D_{IPWAN}=(1-P_m)$$
$$\cdot\frac{S_{mR5}}{W_{avg}}\cdot\left(\frac{(W_{avg}/MSS)\cdot(MSS+TCP_{oh}+IP_{oh})+(S_{ack}+IP_{oh})}{B_{IPWAN}}+d_l\right)\quad(4)$$

In Equation (4), W_{avg} is the mean size of TCP sending window. TCP_{oh} is the size of TCP packet head. IP_{oh} is the size of IP packet head. S_{ack} is the size of acknowledgement packet. B_{IPWAN} is the link bandwidth of IP WAN. d_l is the round-trip link transmission delay.

The service demand D_{Nsn} of network transfer nodes at storage servers is represented:

$$D_{Nsn}=\frac{1-P_m}{STP_m}\cdot IP_{snnum}\cdot\frac{Frames_e}{TRate_e}\quad(5)$$

The service demand D_d of disk I/O nodes is denoted:

$$D_d=\frac{1-P_m}{STP_m}\cdot(1-P_{sn})$$
$$\cdot\left(seek+latency+\frac{STU_m}{TRate_d}\right)\cdot\frac{S_{sn}}{STU_m}\quad(6)$$

In Equation (6), P_{sn} is cache hit probability at storage servers for read operation; $1-P_{sn}$ is the probability of a job accessing disks for read operation. Since the file is far more than system cache size, cache hit probability for write operation can be considered zero. The parameters *seek* and *latency* denote average seek time and average latency respectively. $TRate_d$ is the maximum sustained disk transfer rate.

Table 1. Model configurations

Parameters	Values	Parameters	Values	Parameters	Values	Parameters	Values
S_m	200MB	T_{snpro}	0.028ms	TCP_{oh}	*20B*	P_{sn}	*0.19*
STU_m	16KB	T_{snp}	0.025ms	S_{ack}	*60B*	P_m	*0.13*
STP_m	3	$TRate_e$	100Mb/s	B_{IPWAN}	*100Mb/s*	*Seek*	0.0049s
T_{mpro}	0.027ms	$Frames_e$	1518B	IP_{oh}	*20B*	*latency*	0.00299s
T_{mp}	0.036ms	$TRate_d$	35MB/s	d_l	*0.0005s*	W_{avg}	*32KB*

Table 1 shows the input parameters of the queueing networks model stated above. The size of stripe unit and the number of stripes are both practical configurations of RAID 5 in the NS-RAID system. CPU process time, cache hit probability are the average of practical testing values. NIC and disk parameters are from the datasheets of D-Link DFE530TX and IBM DDYS-T36950.

3.3 Performance bounds analysis

In this section, we will make quantitative analysis for performance bounds of IP-SWAN based on the CQNM, using BJB (Balanced Job Bounds) method [8]. Assume that IP-SWAN model consists of arbitrarily connected K nodes. D_i denotes the service demand of the ith node, where $i \in \{1,...,K\}$, $D_{max}=max\{D_i,\ i \in (1,...,K)\}$, $D_{sum}=\sum_{i=1}^{K} D_i\ ,i \in (1,...,K)$,

$D_{avg}=D_{sum}/K$.

Then the throughput bound of IP-SWAN model is denoted:

$$\frac{N \cdot S_m}{D_{sum}+(N-1)D_{max}} \leqslant X(N) \text{ and}$$

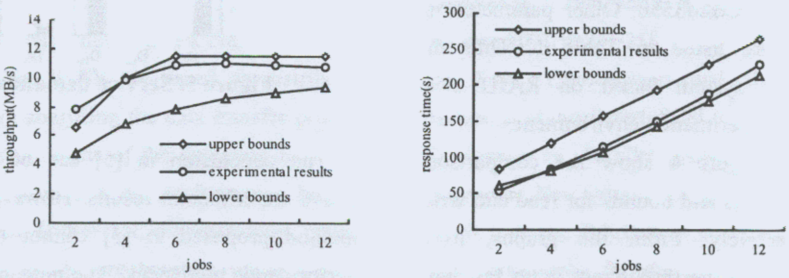
$$X(N) \leqslant min(\frac{S_m}{D_{max}}, \frac{N \cdot S_m}{D_{sum}+(N-1)D_{avg}}) \qquad (7)$$

The I/O response time bound of IP-SWAN model is described:

$$max(ND_{max}, D_{sum}+(N\text{-}1)D_{avg}) \leqslant R(N)$$

$$\text{and } R(N) \leqslant D_{sum}+(N\text{-}1)D_{max} \qquad (8)$$

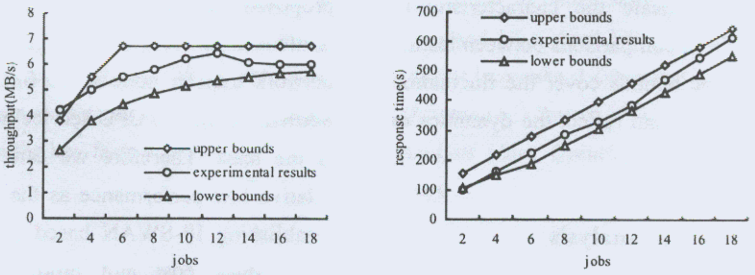
Putting the service demands expressed in Section 3.2 into Equation (7) and (8), we can get the performance analysis model for the throughput and I/O response time of IP-SWAN based on distributed NS-RAID.



(a) Throughput vs. jobs

(b) Response time vs. jobs

Figure 3. Throughput and response time of IP-SWAN for read operation



(a) Throughput vs. jobs

(b) Response time vs. jobs

Figure 4. Throughput and response time of IP-SWAN for write operation

4. Testing and analysis of performance

4.1 Testing results

In this section, we set up an IP-SWAN experimental environment based on distributed NS-RAID to test the performance and then to make comparisons between testing results and the bounds given by performance analysis model proposed in Section 3.3.

The experimental settings of IP-SWAN based on distributed NS-RAID are composed of 5 PCs running on Linux 7.3. One of PCs is configured as router, used for connecting the local and remote network. The Nistnet is installed in the router, used for simulating the IP WAN dynamic performance. The other PCs are used as centre server and storage servers respectively. The configuration of PC is AMD600 CPU, 64M memory, 10/100M D-Link DFE530TX NIC, 36GB IBM DDYS-T36950 SCSI disk. The switch is Cisco3550. Other parameters in this experiment are listed in Table 1. Thus the NS-RAID storage system based on RAID 5 is established in the experimental environment.

Figure 3 and Figure 4 show the comparisons between testing results and bounds for read and write operations respectively. From the graphs, it is obviously to find that the throughput is up fast with the number of jobs increasing in light load. Whereas in heavy load, the overall performance stops growing because of the bottleneck performance limitation. The experimental results indicate the characteristic of slight fluctuations. From comparisons between testing results and bounds, the bounds cover the fluctuation interval of testing results and reflect the dynamics of IP-SWAN performance.

4.2 Performance factors analysis

From Equation (7) and (8), the maximum performance of IP-SWAN depends upon the node with the largest service demand in heavy load. By IP-SWAN performance analysis model, we make computations of service demand at each node and find the node with the largest service demand D_{max} of the service nodes. It is concluded that the node with the largest service demand D_{max} is really the bottleneck node of the system performance. Figure 5 shows the service demands when IP-SWAN conducts read or write operation separately. For read operation, we get to know that IP WAN transfer nodes have the biggest service demand from Figure 5, $D_{max}=17.8$; for write operation, the conclusion is the same, i.e. IP WAN transfer nodes have the biggest service demand, $D_{max}=30.6$. Therefore we conclude that IP WAN transfer nodes are the bottleneck of the whole storage system.

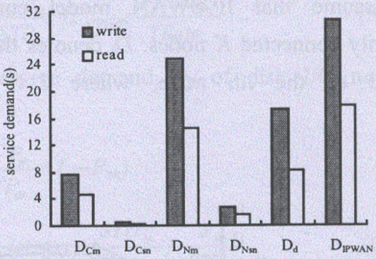


Figure 5. Service demands in IP-SWAN

The conclusion in [5] can be validated by the above experimental results. However, the qualitative method proposed in [5] cannot find the potential performance bottleneck. The potential bottleneck will restrict the maximum system performance when the system bottleneck is improved or the bottleneck performance fluctuates strongly. The model we proposed can be used for analyzing the potential bottleneck as well. From Figure 5, we find that the network transfer node is the first important potential bottleneck and the CPU service node at storage server is the least. Therefore we can use the server with relative low performance as the storage server when establishing IP-SWAN based on NS-RAID. Thus it can reduce cost and produce no influences on performance.

From the above discussion, the node with the biggest service demand, called bottleneck, restricts the overall system performance. Hence the factors affecting on the service demand of bottleneck are factors affecting on the maximum system performance. These factors are called key performance factors. According to Equation (4), cache hit probability, TCP/IP parameters, link bandwidth, transmission delay are key performance factors. The factors affecting on the service demand of other nodes (excluding bottleneck) are called non-key performance factors, which include the processing power of centre server and storage server, disk performance and so on. Based on distinguishing key and non-key performance factors, the optimization of system performance can be simplified through neglecting the non-key factors and focusing on the key factors.

5. Conclusions

To sum up, we have developed a closed queueing networks model for IP-SWAN based on distributed NS-RAID through analyzing the data transfer process of IP-SWAN. At the basis of the queueing networks model, we have proposed the performance bounds analysis model of IP-SWAN. Experimental results show that testing results are between upper bounds and lower bounds, and that the bounds reflect the dynamic trend of the system performance. The bottleneck and the potential bottleneck can be found from the model. Meanwhile the model provides us the criteria for distinguishing the key performance factors and non-key performance factors. When optimizing the system performance, we can neglect the non-key factors and focus on the key factors.

Acknowledgements

This paper is sponsored by National Natural Science Foundation of China (No. 60273031) and Education Ministry Doctoral Research Foundation of China (No. 20020055021).

References

[1] F.Chang, M.Ji, S.-T.A. Leung, *et al.* "Myriad: cost-effective disaster tolerance", *Conference on File and Storage Technologies*. Monterey, CA, 2002, pp103-116.

[2] B.Jones. "The data grid architecture", *Technical Report DataGrid-12-D12.4-33671-3-0*, EUDG, 2002.

[3] Y.Lu and D.H.C.Du, "Performance study of iscsi-based storage subsystems", *IEEE communications magazine*, 2003, 41(8), pp76-82.

[4] P.B.T.Khoo and W.Y.H.Wang, "Introducing a flexible data transport protocol for network storage applications", *10th NASA Mass Storage Systems and Technologies Conference / 19th IEEE Symposium on Mass Storage Systems*, 2002, pp241-258.

[5] X.He, P.Beedanagari and D.Zhou, "Performance evaluation of distributed iSCSI RAID", *12th International Conference on Parallel Architectures and Compilation Techniques*, New Orleans, LA, USA, Sept. 28, 2003.

[6] W.Teck Ng, B.K.Hillyer, E.Shriver, *et al.*, "Obtaining high performance for storage outsourcing", *FAST 2002*, 2002, pp145-158.

[7] Y.-L.Zhu, S.-Y.Zhu and H.Xiong, "Performance analysis and testing of the storage area network", *19th IEEE Symposium on Mass Storage Systems and Technologies*, Maryland, USA, April 2002.

[8] E.D.Lazowska, J.Zahorjan, G.S.Graham and K.C.Sevcik, *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*, Prentice-Hall, 1984