Towards a Performance Model for Virtualised Multi-Tier Storage Systems*

Abigail S. Lebrecht, Peter G. Harrison and William J. Knottenbelt
Department of Computing, Imperial College London

Aims

Storage systems today are expected to deliver consistently high quality of service, despite facing constant demands to store increasingly large quantities of data. The technology pull to massive storage systems has led to the development and widespread adoption of virtualised storage infrastructures that incorporate intelligent storage fabrics. The physical resources underlying such systems are organised into storage tiers, each of which delivers a different cost per capacity ratio. An effective performance model is vital to ensure performance and reliability demands of these systems can be fulfilled. Such a model should abstract the physical features of the underlying disk drives and disk arrays, reflecting faithfully the structure of a virtualised storage system.

Typically, a storage tier will consist of multiple RAID subsystems. Our approach is to initially develop a performance model for an individual disk drive. This can be extended to a RAID model, for various RAID levels, using existing or enhanced techniques. We are specifically looking at RAID levels 0, 0-1, 5 and 6, which are most commonly used. Our goal is to develop a hierarchical, multi-class queueing network performance model of the storage tier and virtualised storage system from these initial model components.

Quality of service constraints can then be optimised at no extra cost by applying discerning device selection and data placement strategies across the tiers. We hope to autonomously and transparently migrate data across tiers and organise data within tiers according to performance benefits. This will contribute towards an overall project ambition of attaining near-optimal performance and reliability over the data lifecycle.

The Model

In initial work, we have developed analytical models of a single zoned disk drive and RAID array using a single server M/G/1 queue and queueing networks respectively. We define service time distributions for the disk drive and model RAID as a fork-join queue [2]. Response time probability distributions calculated from the disk drive and disk array analytical models are validated against device measurements gathered from an Infortrend A16F-G2430 RAID system using 16 Seagate Barracuda ES ST3500630NS (500GB SATA nearline) disks.

Disk Model

Each disk is modelled as an M/G/1 queue. The service time of a disk is modelled as the sum of the rotational latency, seek time and data transfer time [1]. The disks are the slowest part of a disk array; therefore it is fundamentally important to accurately model the disk service time. Consequently, it is important that the effects of disk zoning are incorporated into the service time distribution functions. The data transfer time probability distribution must demonstrate that single sector data transfer time increases as its cylinder gets closer to the disk's circumference. Similarly, the seek time probability distribution should recognise that there is a higher probability of seeking to an outer cylinder, as there are more sectors on those cylinders [3]. Figure 1 compares the cumulative distribution function of our analytical model for data transfer time with device measurements from a ST3500630NS disk.

RAID Model

Analytical results for RAID are generally based on modelling the system as a fork-join queue. However, to date, there have been no exact solutions to calculate the mean response time or response time distribution of a fork-join queue consisting of M/G/1 queues. Thus, we approximate the response time distribution of the fork-join queue, by comparing it to the similar split-merge queue. It is possible to calculate the response time distribution of the split-merge queue exactly using the maximum order statistic [2]. The result from the fork-join queue must be adjusted to make it applicable for RAID. In particular, it must account for parity pre-reads in RAID 5 and 6, mirroring in RAID 0-1 and striping. Figure 2 compares the cumulative distribution of device measurements from the Infortrend A16F-G2430 against the analytical RAID performance model we have developed, for 8 block write requests to RAID 0-1. The array has a stripe width of 128 kB and write caching is disabled for these tests.

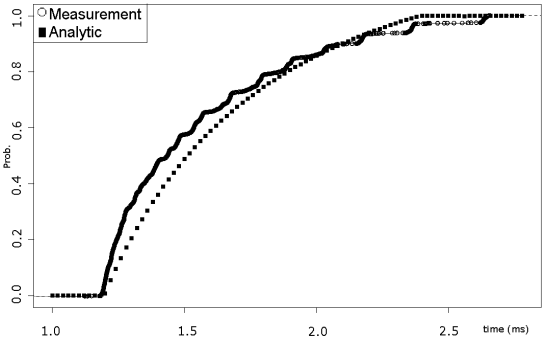

Figure 1: Data transfer time of 100 MB requests on a Seagate ST3500630NS disk, compared with analytical model of zoned data transfer time

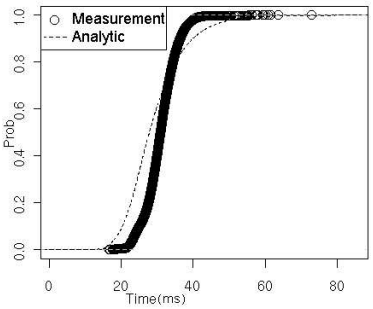

Figure 2: Response time cumulative distribution for 8x128 kB write requests on Infortrend A16F-G2430 RAID 0-1 with 16 Seagate ST3500630NS disks against the analytical RAID model

Future Work

Initial results suggest that the disk drive queueing model is sufficiently accurate. The RAID performance model needs refinements to take factors into account that are not presently effectively depicted, including caching. With improved accuracy from the RAID performance model, we expect to incorporate it into a performance model of a storage tier and extend that into a model for a hierarchical multi-tier virtualised system.

References

[1] S. Chen and D. Towsley. A performance evaluation of RAID architectures. *IEEE Transactions on Computers*, 45(10):1116–1130, 1996.

[2] A. S. Lebrecht and W. J. Knottenbelt. Response time approximations in fork-join queues. In *23rd Annual UK Performance Engineering Workshop (UKPEW)*, July 2007.

[3] S. Zertal and P. G. Harrison. Multi-RAID queueing model with zoned disks. In *High Performance Computing and Simulation Conference (HPCS'07)*, June 2007.

*This work is supported by Engineering and Physical Sciences Research Council grant number EP/F010192/1