Work in Progress: On the Scalability of Storage Sub-system Back-end Network

Work in Progress: On The Scalability of Storage Sub-System Back-end Network

Yan Li[†], Tim Courtney[‡], Roland N. Ibbett[†], Nigel Topham[†]

[†]Institute for Computing Systems Architecture,
School of Informatics, University of Edinburgh
Y.Li-24@sms.ed.ac.uk, {rni, npt}@inf.ed.ac.uk

[‡] Xyratex, 1000-80 Langstone Technology Park,
Langstone Road, Havant, Hampshire, PO9 1SA, UK
tim_courtney@xyratex.com

The aim of this on-going work is to study the scalability of the back-end network of storage sub-systems in terms of the number of disks that can be linked to the network. It is well known that without considering the limitation of back-end network, increasing the number of disks in a RAID based storage system will increase the parallelism, and so can lead to a higher performance. Moreover, to save money on the back-end network, it is common practice to scale the number of disks rather than the number of independent access pathways. However, in a real system there is a limitation on the scale of storage sub-systems (controller cache size and number of disks that can be included in one system) due to the limitation of interconnection network. This is because the back-end interconnection networks are shared by all the disks and the RAID controllers in a storage sub-system. The more disks are added to the system, the higher the contention for the shared media. When the number of disks and cache size in a RAID system reaches a certain threshold, there will be no further gain in performance by adding more disk or cache due to the saturation of the back-end network. Therefore, in order to design a scalable storage sub-system it is critical to study the saturation characteristics and scalability of the back-end network. Previous work has focussed on sequential accesses only when working out when the back-end network becomes saturated, this does not represent a 'normal' workload. This work uses a workload based on the Storage Performance Council SPC-1 benchmark and so uses a representative loading.

Our research has chosen Fibre Channel (FC) Switched Bunch of Disks (SBOD) [1] as the research subject since this represents the current state of art in scalable back-end storage sub-system solutions.

The research include two parts. First, a simplified theoretical model has been developed to study to the scalability of FC SBOD, which confirms that the number of disks that saturate the back-end network is mainly decided by the characteristic of the workload, the stripe size, RAID protection level (RAID 5 or RAID 6) and the network bandwidth. Second, detailed simulations were carried out to studied the scalability of FC SBOD in a more accurate way. A discrete-event driven simulation model called SIMRAID was developed to model the storage sub-system. Unlike most storage system simulators which ignore the effect of the back-end network, SIMRAID explicitly simulates the FC network so that it is capable of studying the performance of the network. Moreover, in contrast to some FC network simulators which simulate FC at word/frame level, SIMRAID simulates the FC-AL at a higher abstraction level coupled with topology knowledge coded into the simulation to enhance accuracy, resulting a much faster simulation speed than the word/frame level simulation at the expense of a slight accuracy reduction. The accuracy of SIMRAID has been verified through simulation of a system for which there are published SPC-1 benchmark results. This simulation shows a maximum inaccuracy of 4% coupled to a speed-up in simulation time of over 5 orders of magnitude thus validating the simulation approach. The simulation includes two stages of study. First, simulations are carried out to study the scalability of FC in a storage sub-system without any cache. We first study the number of disks that saturate a $2Gbps$ FC SBOD by using one SBOD under RAID5 and RAID6 protection levels using an SPC-1 based workload. The simulation gives a result of 48 disks for RAID5 and 53 for RAID6 when the stripe size is $16KB$. Further simulations will then be carried out to contrast using 4Gbps FC with and dual $2Gbps$ FC port. The second part of the simulation is to study the scalability of FC SBOD in a cached system. We will first study relationship between cache size and the number of disks by setting the network bandwidth to near infinite so that it will not saturate. This tells us the bandwidth required to obtain maximum performance for a given cache size and number of disks. We will also know the combination of number of disks and cache size that will saturate a given network bandwidth. Last, we will study the network bandwidth requirement of the cache coherency.

REFERENCES

[1] Emulex. [TM]InSpeed soc 320 embedded storage switch, March 2003. available at http://www.emulex.com/products/embeddedswitch/320/ds.pdf.