On The Scalability of Storage Sub-System Back-end Network

Yan Li, Roland Ibbett, Nigel Topham and Tim Courtney[‡]

School of Informatics
University of Edinburgh
[‡]Xyratex, UK

Project Motivation

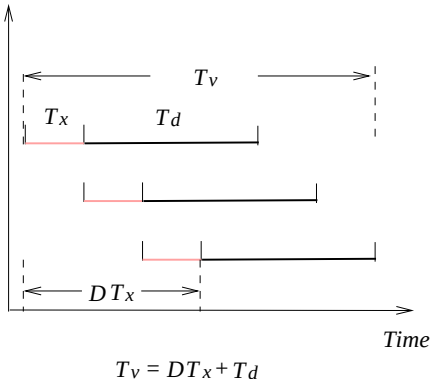
- Theoretically the more disks used in a disk array, the higher the degree of parallesim, so leading to larger the performance potential performance benefits.
- However, in a real system there is a limitation on the scale of RAID systems due to the limitation of interconnection network.
- The more disks 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.

Project Goal

- Investigate the capacity of interconnection networks in terms of the numbers of disks that can be included in one chain. In particular, Fibre Channel (FC) SBOD is chosen as the interconnection network.
  - Give a certain number of disks and cache size, how much bandwidth is necessary to support them to get the maximum performance?
  - Likewise, how many disks (and cache) can be connected to a 2GFC (4GFC) port?

- Assumptions:
  - The request size of random access workload is equal to the size of stripe unit.
  - The access address of each request is aligned to the stripe unit boundary.
  - The capacity of the RAID system keeps fixed for all the study.
  - The queue length of disk is 1, ie., no disk command waits in the disk for service.
  - The FC SBOD is chosen as the research subject.
- Disk command transmission time $Tx = \frac{S}{B} + overhead.$.
- Disk command execution time $Td = \frac{S}{B_{port}} + \frac{S}{B_{media}} + T_{seek}$.
- S size of stripe unit; B network bandwidth;

Large Sequential Access

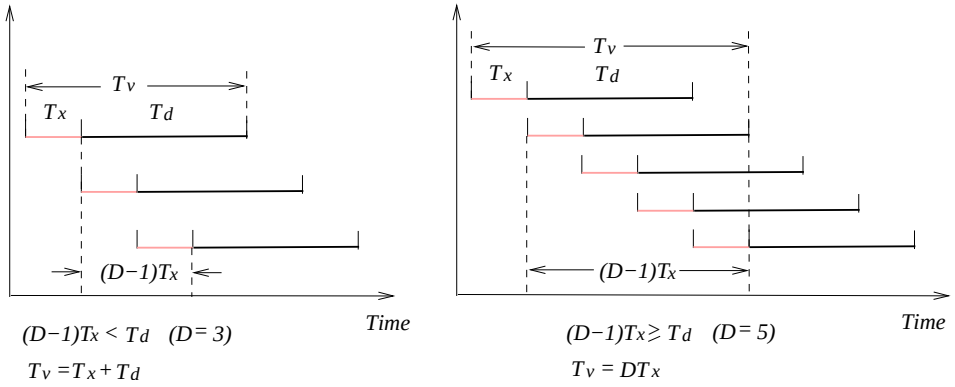


- For sequential access workload, a large volume command is divided into D disk commands, so that the response time for that command $Tv = DTx + Td$.

- Throughput is the major performance metric:

$$\textbf{Throughput} = \frac{K}{Tv} = \frac{K}{DTx + Td}.$$

Small Random Access

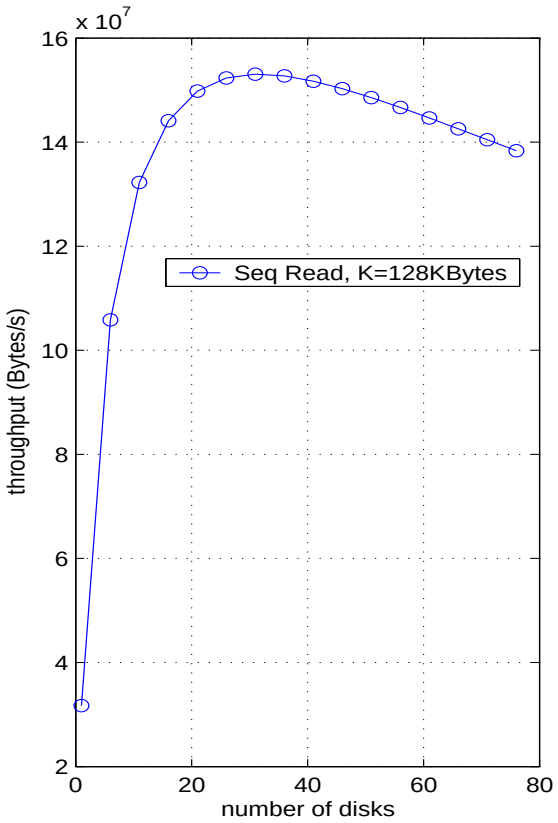


$$\mathbf{Tv} = \begin{cases} Tx + Td & (D-1) * Tx < Td \\ D * Tx & (D-1) * Tx >= Td \end{cases}$$

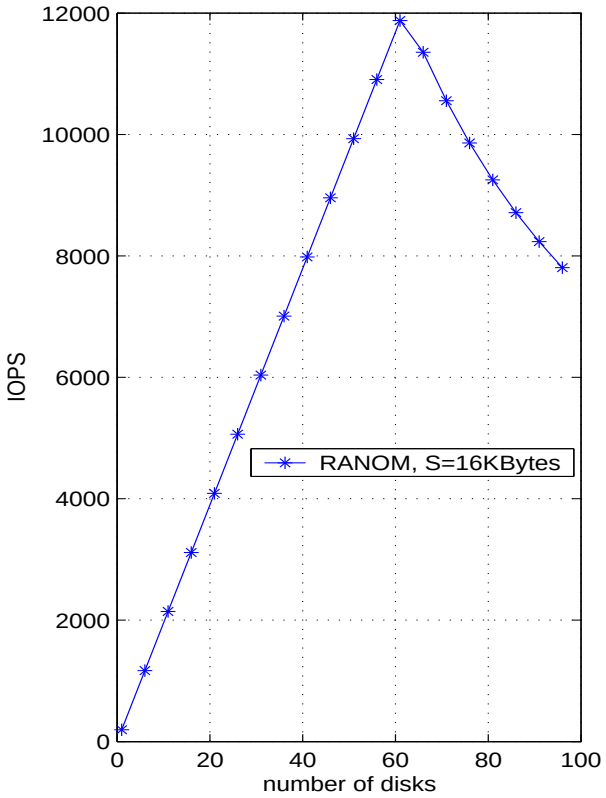
- For random access workload, IOPS is the major performance metric,

$$\mathbf{IOPS} = \begin{cases} \dfrac{D}{Tx + Td} & (D-1) * Tx < Td \\ \dfrac{Td}{Tx} * \dfrac{1}{D * Tx} & (D-1) * Tx >= Td \end{cases}$$

Analytical Results (no cache)



Sequential Access



Random Access

General Model

- $B = F(D, C, S, L, P) = Num_d * S + overhead$

  B, the bandwidth required to achieve the maximum performance with D disks and C cache in system.

  D: number of disks

  C: cache size

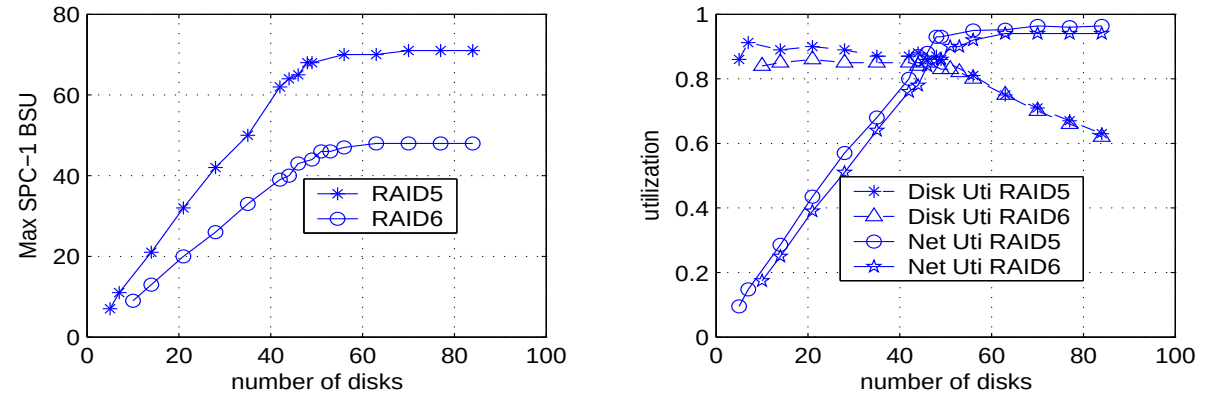
  S: size of stripe unit

  L: workload characteristic

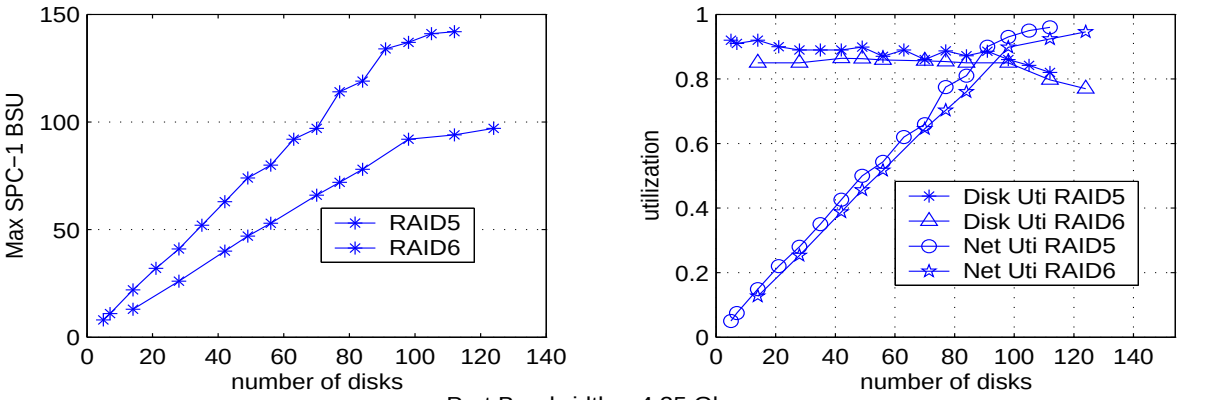
  P: cache destage threshold, P=0 for our study

  Num_d: number of disk commands send to disk per second

Systems without Cache ($S = 16KB$)

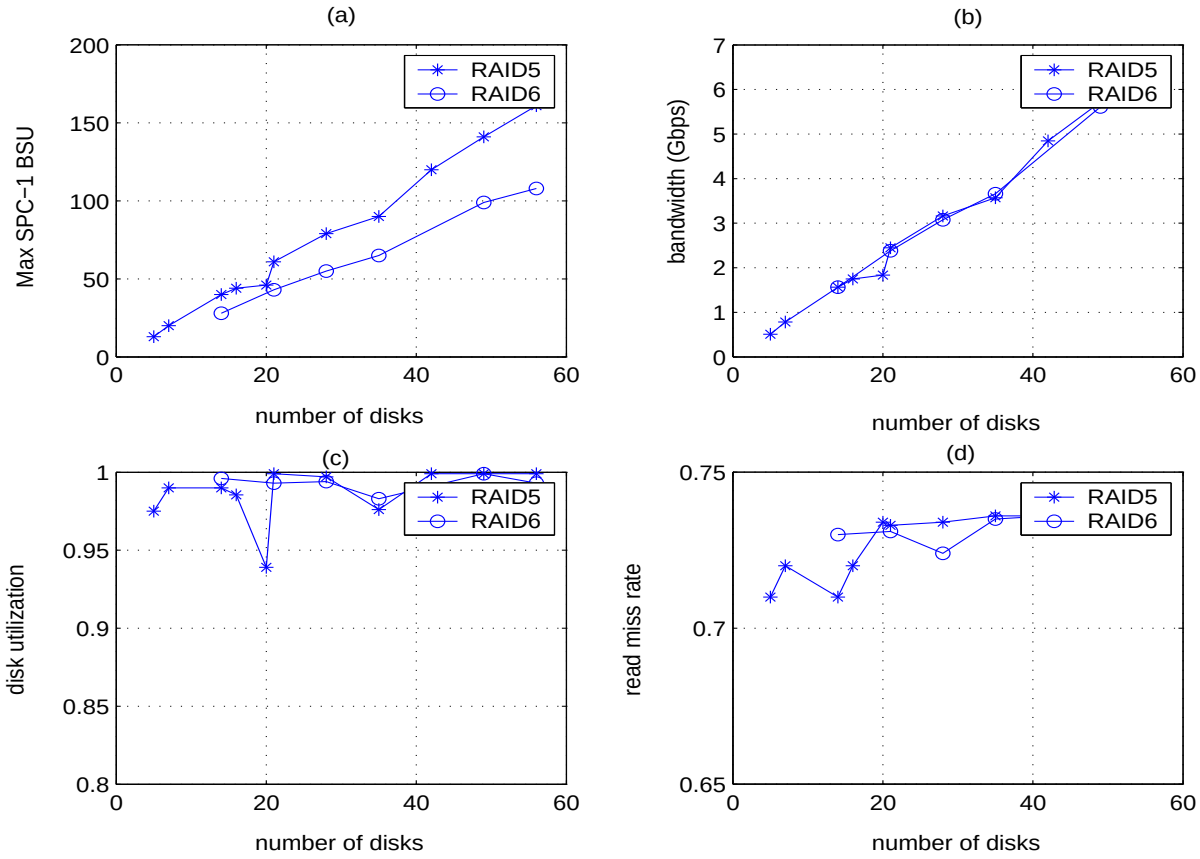


Port Bandwidth = 2.125 Gbps



Port Bandwidth = 4.25 Gbps

Systems with Cache ($S = 32KB$)



Summary

- When this is no cache, a 2G FC port is able to support up to 46 disks for RAID5 and 53 disks for RAID6 (size of stripe unit = 16kBytes).
- With enough cache, a 2G FC port is able to support up to 18 disks under OLTP like workload. (size of stripe unit = 32KBytes).
- When there is enough cache, the bandwidth required to support a certain number of disks is fixed. It is irrelevant with protection level and cache size.

Future Work

- Study the scalability of back-end network when the size of stripe unit is 16k and the system performance.

- Study the network bandwidth requirement when there is cache coherency between two controllers.