Fun with Numbers: Storage Provisioning

Catharine van Ingen

December 2005

Technical Report
MSR-TR-2005-167

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Fun with Numbers: Storage Provisioning

Catharine van Ingen
vaningen@microsoft.com
Microsoft Research, 455 Market St., Suite 1690, San Francisco, CA 94105
http://research.microsoft.com/barc

Executive summary: *Storage hardware costs have often been quoted as price per capacity ($/GB). A better metric is total cost per user as deployed. This balances performance, capacity and packaging constraints. That being said, it is important to recognize that hardware is only one part of the total cost of ownership for storage. Storage is now cheap and people remain expensive.*

Storage hardware vendors commonly tout storage costs in dollars per GB. But what does that really mean when the storage is actually deployed at a customer site? The actual cost must include the extra storage purchased for performance, availability, and manageability. This is particularly true when the amount of storage that must be purchased is determined more by performance (IO operations or IOPs) rather than capacity (GB) or when significant additional hardware is purchased for availability.

A simple model was constructed to investigate this. The model begins with a simple model for storage hardware and applies an abstracted application workload. A detailed example is presented to illustrate the use of the model. Other workloads and other storage hardware can be used with this methodology.

The Model

The storage hardware system was modeled as a drawer. The drawer contains a number of disks and is depreciated over a specified lifetime. Failed disks are replaced. This approach allows enterprise (SCSI) disks to be simply compared with desktop (SATA) disks. For the example used in this report, we used fifteen disks in the drawer and a five year lifetime.[1]

Eight different currently shipping disk drives were considered as shown in Table 1. All prices were obtained from an internet price comparison shopping site; the lowest non-OEM unit price for a new unit was used. Note that the cost for the storage controller and interconnect are not included in this pricing. The associated pricing for enterprise drives can be significant and on the order of the price of the desktop drives in the drawer.

The enterprise disks offer more performance and reliability at a price premium. The desktop drives offer significant capacity at very low price. The 60-90% higher performance advantage of the enterprise disks is due to both higher RPM as well as somewhat smaller random seek time.

The nominal quoted MTBF of enterprise disks is about 1.5M power-on hours or over twice that of the nominal quoted MTBF of desktop disks of about 600K hours. For this note, drive failure rates of 1% per year for the enterprise disks and 7% per year for the desktop disks were used. These numbers are compromises between the very much lower quoted rates and anecdotal yet empirically valid data from anonymous MSN properties and internet sites. For the five year depreciation period, these failure rates correspond to one enterprise disk failure and six desktop disk failures.

The storage model includes 2-way mirrors, 3-way mirrors and 4-way RAID5 configurations. Each configuration was modeled with 10% reserved free space; administrators commonly reserve between 10% and 30% free space on a volume for temporary space allocation during storage management operations such as defragmentation or object recovery (eg recovering a file, mailbox, or database table). Each configuration was also modeled with and

[1] The number of disks was chosen based on currently shipping configurations. Depreciating this hardware over 3 to 5 years is common practice.

Table 1. Disk Drive Properties								
	Enterprise				Desktop			
Capacity (GB)	73.00	73.00	146.00	300.00	80.00	120.00	250.00	400.00
RPM	10000	15000	15000	10000	7200	7200	7200	7200
Random IOPs	130.00	150.00	150.00	130.00	80.00	80.00	80.00	80.00
Failure per year	0.01	0.01	0.01	0.01	0.07	0.07	0.07	0.07
$ per GB	2.59	5.45	5.47	2.31	0.65	0.58	0.52	0.76
$ per disk	189	398	799	694	52	69	130	304
$ per drawer	2835	5970	11985	10410	780	1035	1950	4560

without 10% reserved for copy-on-write snapshots.[2]

The workload was modeled simply by heat and read-to-write ratio. Heat is the number of IOPs per GB applied by the application and captures how hard the application loads the disk drives. The read-to-write ratio is important because reads and writes cause will different number of actual disk IOPs due to both RAID and snapshots.

The worked example in this note uses a moderately heavy write workload with read-to-write ratio of 2.5 to 1 and .8 IOP per user.[3] Two different end user allocations of 100 MB and 2 GB were considered corresponding to a heat of 8 and .4 IOP per GB.

RAID, snapshots, and free space impose capacity and IOP taxes. Table 2 summarizes the percentage of the raw disk capacity that is available to an application after subtracting the RAID, snapshot, and free space overheads for the example workload. As expected, the RAID5 configuration offers the most capacity while the 3-way mirror offers the least. Also as expected, the RAID5 configuration imposes the highest IOP tax and copy-on-write snapshots are a significant tax. Because this is a random access workload, there is little locality of reference and any cache efficiency will be low. Note that the taxes in the table above are for healthy RAID sets; during repair, the tax is higher as additional IOPs are needed to reconstruct the failed mirror or RAID 5 set member and compute the data from the RAID 5 parity.

Table 2. RAID and snapshot taxes			
	2-way mirror	3-way mirror	RAID5
Raw capacity available for data (%)	45	30	68
IOP per user IOP	2.14	2.71	3.26
IOP per user IOP without snapshot	1.29	1.57	1.86

100 MB user allocation results

The results for 100 MB user allocations without snapshots are shown in Figure 1. All of the configurations are significantly IOPs performance limited.

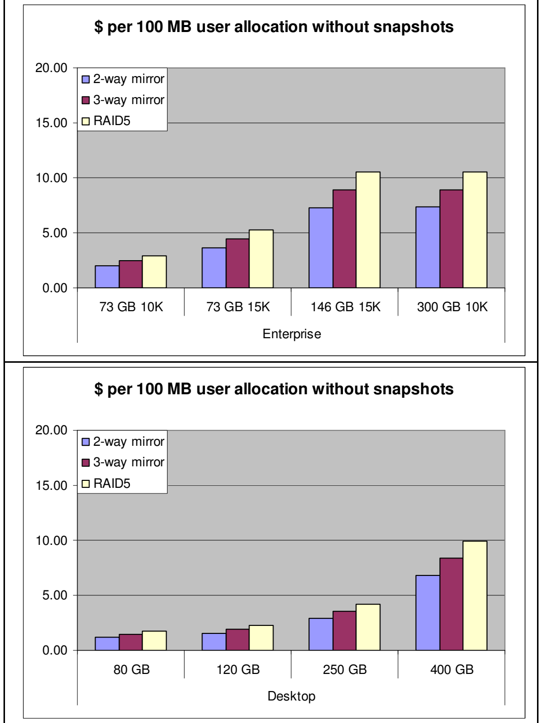
Only 27% of the capacity surface of the 2-way enterprise mirror is occupied by user data with the 73 GB enterprise disk and only 4% is occupied with the 400 GB desktop disk. Another way of looking at this is that the 73 GB drawer could support user allocations of about 400 MB with no change in hardware. Note that these numbers are after the capacity taxes have been applied; the total user data is about 14% of the

[2] Note also that when copy-on-write is used, multiple points in time (snapshot layers) can be present with the same overhead and marginal increase in storage where as when a full mirror is used only one point in time can be present. The single point in time is useful for backup; multiple layers are more useful for online single object recovery such as single file or mailbox recovery.

[3] This workload is based on the current Exchange 11 database. The average peak IO on Microsoft Exchange servers is approximately .8 IOPs per user with 2.5 reads-to-writes. While Microsoft is a heavy mail user, the peak reported for customer sites can be as high as 1.2 IOPs per user. Common Microsoft user allocations are 100 MB; gmail offers 2.5 GB. Also note that future versions of Exchange are may have very different heat and read-to write ratio.

raw capacity of the 73 GB drawer and 2% of the 400 GB drawer.

Figure 1: 100 MB per user allocation without snapshots

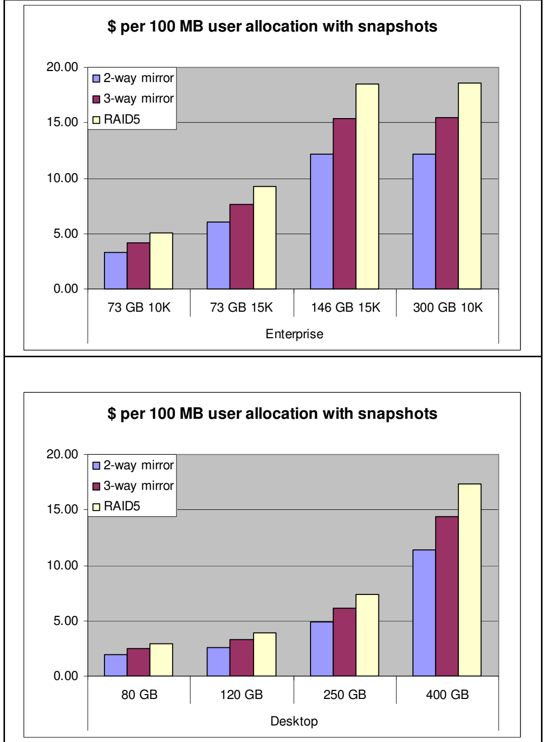


The mirror configurations are less expensive than the RAID5 configurations because the configurations are IOPS limited and RAID5 has a substantial IOPS tax. While the mirror GB tax is higher, the mirror IOP tax is lower. The 2-way and 3-way enterprise mirror configuration are roughly 70% and 85% the cost of the RAID5 configuration (this is just the storage cost, if one included the server, front-end, and network costs, the difference would be less).

The 73 GB enterprise disk configurations are roughly 65% more expensive than the 80 GB desktop disk configurations. The raw cost per GB difference is roughly 400%. The smaller differential is because more desktop disks must be purchased to get the same level of performance and reliability. The enterprise disk drawer supports 1000-1500 users depending on the RAID configuration; the desktop drawer supports only 650-900 users. Six desktop disks must be replaced over the five year lifetime; only one enterprise disk must be replaced.

The results for 100 MB user allocations with snapshots are shown in Figure 2. These configurations are even more performance limited as snapshots add a significant IOP tax to writes. The cost per user is roughly 70% more with snapshots. This is in direct proportion to the increased IOP load cased by the copy-on-write.

Figure 2: 100 MB per user allocation with snapshots comparison



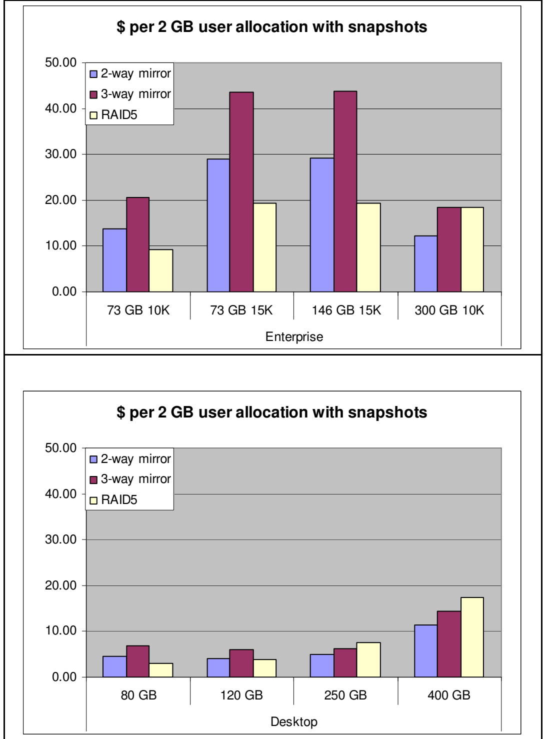
2 GB user allocation results

Figure 3 shows the results for larger 2 GB user allocations with snapshots. These configurations range from performance limited to capacity limited.

Two of the RAID 5 configurations are roughly balanced. These are the 146 GB enterprise disk and 80 GB desktop disk configurations. The enterprise drawer supports almost twice as many users as the desktop drawer (80% more). The cost per user with enterprise disks is about 2.6 times more than the desktop configurations.

The mirror configurations with the 300 GB enterprise disks and the 250 GB desktop disks are also roughly balanced. With 2-way mirrors, the enterprise drawer supports twice the number of user at 2.5 times more cost per user. With 3-

Figure 3: 2 GB per user allocation with snapshots comparison



way mirrors, the enterprise drawer supports only 10% more users at almost 3 times the cost.

The other enterprise disk configurations are capacity limited with the exception of the 300 GB RAID 5 configuration which is performance limited. When capacity is the limitation, RAID 5 offers lower cost per user for a given disk. The three RAID 5 configurations with 73 GB, 146 GB, and 300 GB all offer about the same cost per user.

The least cost enterprise drive configuration is the 2-way mirror with 300 GB disks at $12 per user without snapshots. The additional performance of the 15000 RPM disks is not used. The smaller 73 GB 10000 RPM disk drawer is about a quarter of the cost of the 300 GB drawer, but supports less than a quarter of the number of users.

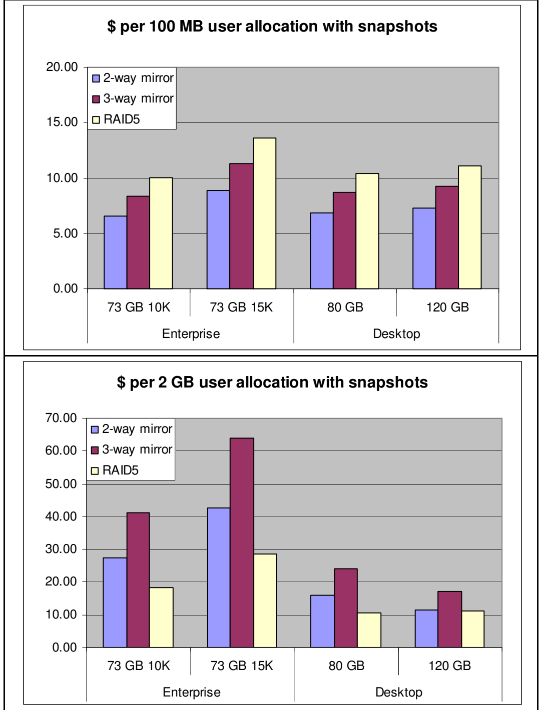
The desktop disk configurations range widely from very performance limited to very capacity limited. The mirror configurations with 80 GB and 120 GB disks are capacity limited. The RAID 5 configuration with 250 GB disks and all configurations with 400 GB disks are performance limited. Only the 400 GB configurations are over $8 per user. With 120 GB disks, the cost of 2-way mirror or RAID 5 is roughly the same.

Other results

Disk cost is only part of the story. The drawer contains additional components such as controllers, power supplies, and fans and must be housed in a rack. Labor is also required to replace disks.

Figure 4 shows the results when a $3000 fixed overhead per drawer is added to the four smaller disk configurations.[4] The fixed overhead taxes the desktop disk configurations more than the enterprise configurations because those configurations support fewer users per drawer. For the 100 MB user allocations, there is little difference between the 73 GB enterprise disks and the 80 GB desktop disks. For the 2 GB allocations, 120 GB disk is now a clearly a more cost effective choice than the 80 GB disk.

Figure 4: Fixed drawer $3000 overhead comparison



[4] The $3000 was selected based on a packaging overhead estimate from an anonymous subsystem OEM and rack mounting cost from an anonymous internet property. The estimate may be somewhat high given that the disk cost to the subsystem manufacturer would be less than off a public web site.

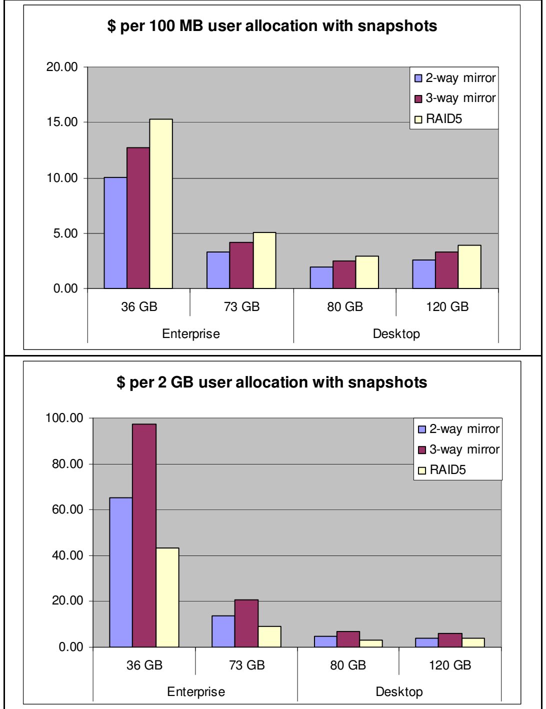
If the cost of the drawer is depreciated over a 3 year rather than 5 year lifetime, the desktop disks become slightly more economical. Fewer desktop drives are replaced over the 3 years reducing the cost of the drawer.

If the drawer contains 20 disks rather than 15, there is very little change in the above. The number of desktop drives that must be replaced increases, but that change is relatively small.

If a less write intensive workload or a more sequential workload was used the differences in the cost per user become smaller. The overhead of a copy-on-write snapshot is magnified by the hardware RAID. The fewer writes, the fewer total IOPs into the drawer. With fewer writes, the balance point between buying for performance and buying for capacity shifts to allow a hotter workload.

A historical comparison is also interesting. Figure 5 compares a 2000 36 GB enterprise disk with the 2005 73 GB 10K enterprise disk and the two smaller desktop disks. The raw cost of the 2000 enterprise disk is 4 times that of the 2005 enterprise disk and almost 20 times that of the desktop drives. With 100 MB user allocations, the 2000 enterprise disk mirror configurations are roughly three times more expensive than the 2005 enterprise drive; the RAID 5 configuration is roughly 5 times more expensive. With 2 GB user allocation, all 2000 enterprise disk configurations are roughly 5 times more expensive. Moreover, the absolute magnitude change of the per user cost is more – a 100 MB user allocation in 2000 cost more than a 2 GB user allocation in 2005.

Figure 5: 36 GB 1995 disk comparison



Summary

Storage hardware is purchased for performance, capacity and availability. RAID imposes both performance and capacity taxes. The least expensive disk or fastest disk may not yield the least expensive solution when deployed. The best solution may well “waste” capacity for performance and availability; that is certainly counter-intuitive.

Of course, this simplistic hardware cost comparison does not tell the full story. People, not hardware, likely dominate the total customer cost of ownership in 2005. Today, the total cost of the drawer is a small part of the cost of administrator, helpdesk, and networking costs. This simple comparison does call into question comparing the commonly quoted cost per raw capacity and the importance in computing the *actual cost per user* even when done simplistically.

Thanks to Chris Mitchell, Bob Fitzgerald, Dan Stevenson, and Ed Tremblay for asking the original questions that motivated this model and reviewing drafts of this note. As always, Jim Gray provided inspiration.