

STORAGE PERFORMANCE MANAGEMENT:

# How Tiered Environments Can

**LOWER COSTS****&****IMPROVE PERFORMANCE**

By Gilbert Houtekamer, Ph.D., &amp; Wim Oudshoorn

Storage system design always involves a trade-off between cost and performance. A storage system full of small, fast disks may provide great performance and throughput, but the benefits don't always justify the costs. "Tiering," or using a mixture of faster and slower disks, is likely to produce a better price-to-performance ratio. You can reserve the fastest, most costly disks for the workloads that really need them and let large disks handle less active workloads.

Tiered configurations are more challenging to design and manage, but with care, they can be significantly less expensive and offer great performance. This article examines how to assess whether tiering can be used in your environment to reduce storage acquisition costs while maintaining the required performance levels. It also

explains how to design a tiered solution with confidence.

#### Potential

When carefully designed, multi-tier solutions can be much more economical. The savings will be greater when there's a mix of highly active workloads with high-performance requirements and less active workloads with less stringent performance requirements. In such situations, a multi-tier storage system may yield a less costly and faster solution than is achievable with a single tier.

Most of today's mainframe installations predominantly use 146GB 15k RPM Fibre Channel (FC) disks. When acquiring new hardware, there are many more options available, including playing it safe with the same drive technol-

ogy or moving everything to larger drives. If you can altogether move to drives that are twice as large, only half the number of drives will be needed, significantly reducing acquisition, maintenance, and environmental costs. Alternatively, if one-third of your data is quite active and needs the 146GB drives, it may be possible to store the remaining data on disks as large as 450GB. This combination of disks also would decrease the number of Redundant Arrays of Independent Disks (RAID) array groups to a little more than half the number of original groups. An original configuration with 36 146GB RAID array groups would have just  $12 + 24/3 = 20$  RAID array groups in such a new configuration.

Of course, one RAID array group of larger capacity disks is often more

expensive than one RAID array group of smaller disks, so halving the number of RAID array groups may not halve the acquisition costs. As prices between larger and smaller disks vary per vendor, we've estimated the relative costs listed in Figure 1 for the examples in this article. Based on the numbers in the relative costs table, the cost difference between the tiered solution with 12 groups of 146GB disks and eight groups of 450GB disks, and the original setup of 36 groups of 146GB disks is 28 compared to 36, a 22 percent reduction in acquisition cost while still using 15k RPM technology. For a more detailed example of the possible cost savings, see the tiering examples in the accompanying sidebar.

### Risks and Drawbacks

The drawback of a storage configuration with multiple tiers is that it's more complex to manage. Once you manage multiple tiers, you need to decide which data needs to be handled by which drive technology. These decisions are crucial for the effective performance of the multi-tier configuration; mistakes can compromise an otherwise good design.

In smaller configurations, the separation of data into tiers makes it more challenging to effectively use drive capacity. If you can't effectively use the space, you might need to configure more RAID arrays, precluding some of the cost savings associated with multi-tier.

A multi-tier configuration is designed with a specific workload in mind. But if the workload access pattern changes or grows more than anticipated, a single-tier solution is easier to extend by simply adding more disks. A multi-tier solution may be less expensive to extend, but you first need to decide which disk type to add for the best result.

### Designing a Multi-Tier Configuration

Designing a storage configuration will always be something of an art, but it's possible to design a multi-tier solu-

Drive type	Relative cost per RAID 5 (7+1) array	Relative cost per GB
146GB 15k	1	2.7
300GB 15k	1.5	2
450GB 15k	2	1.8
400GB 10k	1	1

Figure 1: Assumed Drive Cost

tion with confidence by using the following steps, most of which also apply to non-tiered storage systems:

1. Measure workload characteristics.
2. Split the workload into two parts, one with a higher and one with a lower back-end density.
3. List the space and performance criteria for each of the two workload parts.
4. List the capacities and performance capabilities for each viable disk and RAID type combination.
5. Combine the above information to find the number of RAID arrays required for each disk/RAID type for the two workload parts. Pick the most attractive of the options.
6. Assess feasibility of this option given the individual volume characteristics. If not feasible, return to step two or three.

Let's consider each of the steps in more detail:

**1. Measuring the workload:** This is a crucial step. Accurate, complete information about your workload will help you design a multi-tier configuration with confidence. When less information is available, all the subsequent steps also will be less accurate and risks will be greater.

For designing any new disk configuration, the following variables are indispensable:

- Back-end I/O rate for the chosen RAID scheme
- Size of the data in GB or cylinders.

These numbers are required per storage group or application group to create a good workload split in step 2. The information per volume is required to create correct volume mapping in step 6.

Because the access pattern doesn't remain constant all day, this data is required for more than one point in time. The workload profile during the batch window will be quite different from the online period, and the configuration must clearly support both. Also, many installations run week-end (or month-end) workloads that are different again. A minimum of one week of data should be used for tiering studies.

### Note on Back-End I/O Rate

The back-end I/O rate isn't the same as the number of host-issued I/Os.

Rather, it's the number of I/Os issued to the physical drive, resulting indirectly, from host-issued I/Os. The actual number depends on the type of I/Os, the effect of storage system cache, and the RAID scheme used. This number is crucial for determining whether a specific drive technology can handle the workload; the number directly determines the drive utilization.

### How to Obtain This Data

Detailed workload information can be extracted from Resource Measurement Facility (RMF) in z/OS systems or by using software products that will create and present all the variables discussed in ready-to-use tables and charts. Alternatively, you can write reporting product programs to process and enhance the RMF information by computing the back-end I/O rate from cache and device statistics.

**2. Splitting the workload:** The hardest part is to make the right workload split. The measurement numbers you obtained from the analysis described earlier will help you create groups of workloads with high vs. low back-end access density. You should ensure the high-access density part isn't so back-end-intensive that it wouldn't be possible for even a fast disk type to support the access density. Examining workload profiles per storage group will usually give you a good starting point for creating an optimal split.

Manually performing this workload split is tedious and error-prone especially because you'll need to consider multiple periods (batch, online) as discussed. Specialized software can automate most of the work involved. Ideally, your software will create all possible workload splits and run them through the remaining steps, finding the best option for you.

**3. Determining space and performance criteria:** For the space criterion, this step is easy. You simply need to know how much capacity needs to be allocated to contain the data sets and also allow for some growth. The performance criterion has two aspects: the disk response time and the disk busy. For your final choice of configuration, response time should be as low as possible and the disk busy as high as possible without impacting performance.

For planning purposes, you need to establish the targets for maximum acceptable disk busy levels and maxi-

imum acceptable back-end disk response time. The maximum disk busy levels determine how resilient the design will be against unexpected spikes in activity. If the target is under 50 percent, doubling the I/O load (as compared to the current situation) would still work, albeit at high response times. If the target is higher than 50 percent, there's less elbow room for unexpected spikes and workload growth.

In a healthy, well-managed storage solution, the disks shouldn't be more than 50 percent busy during the online period. If the disk busy level exceeds 50 percent, there'll be significant queuing on the disk devices; this will rapidly increase the read miss response time—to a point where it becomes unacceptable for an online application. During the batch window, higher disk drive busy levels can occur because batch jobs

will typically try to get as many I/Os as possible per second, automatically driving up the busy percentage of the drive. As long as other workloads don't suffer, that isn't a problem.

The maximum back-end disk response time criterion is important; you must ensure that the worst-case performance is acceptable. As the workload intensity on the back-end disk drives (disk busy) increases, disk response time will increase.

The lower the performance targets, the more costly the solution becomes. Also, the minimum possible random disk access time for a 15k drive is around 6 ms, so unless you consider Solid State Drives (SSDs), you won't be able to get a better response time than that. Furthermore, hard disks can perform only at 6 ms response time levels when the I/O rates are quite low, so if you actually set the response time target to 6 ms, the cost of the solution will become high because of the extra drives you'd need to deploy. Appropriate values for response time targets range from 10 to 15 ms. Base your choice of targets on business requirements and cost levels.

These disk response times are different from the response times your applications experience. Application I/Os will have to wait only for the disk to service the I/O if the I/O is a read miss, normally a minority of the I/Os. For instance, for a read/write ratio of one (50 percent reads) and a read hit percentage of 90 percent, only one in every 20 I/Os will actually have to wait for the

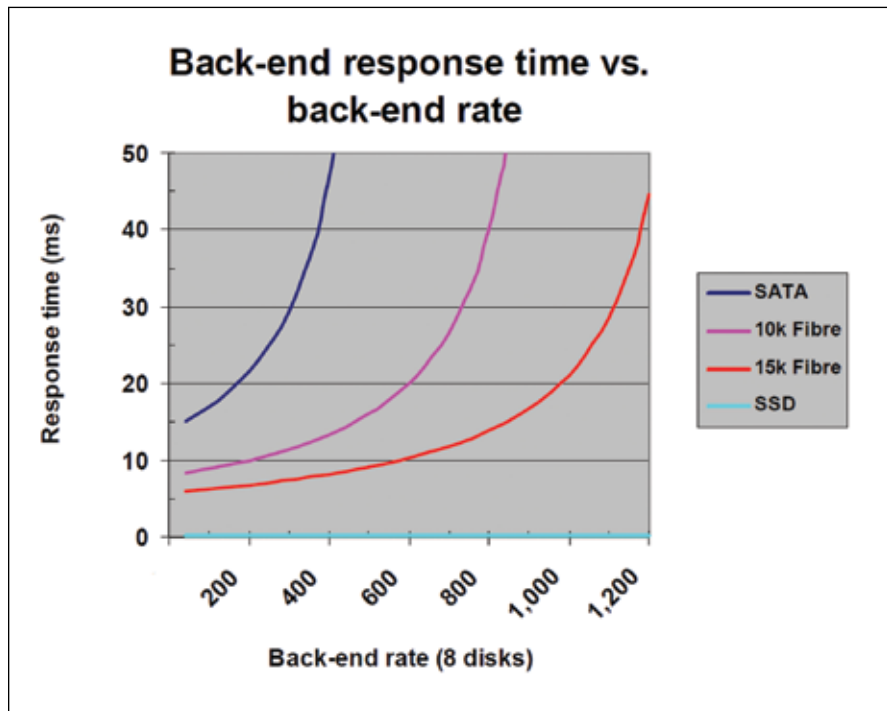


Figure 2: Response Time as Function of the Back-End Rate for Different Disk Technologies

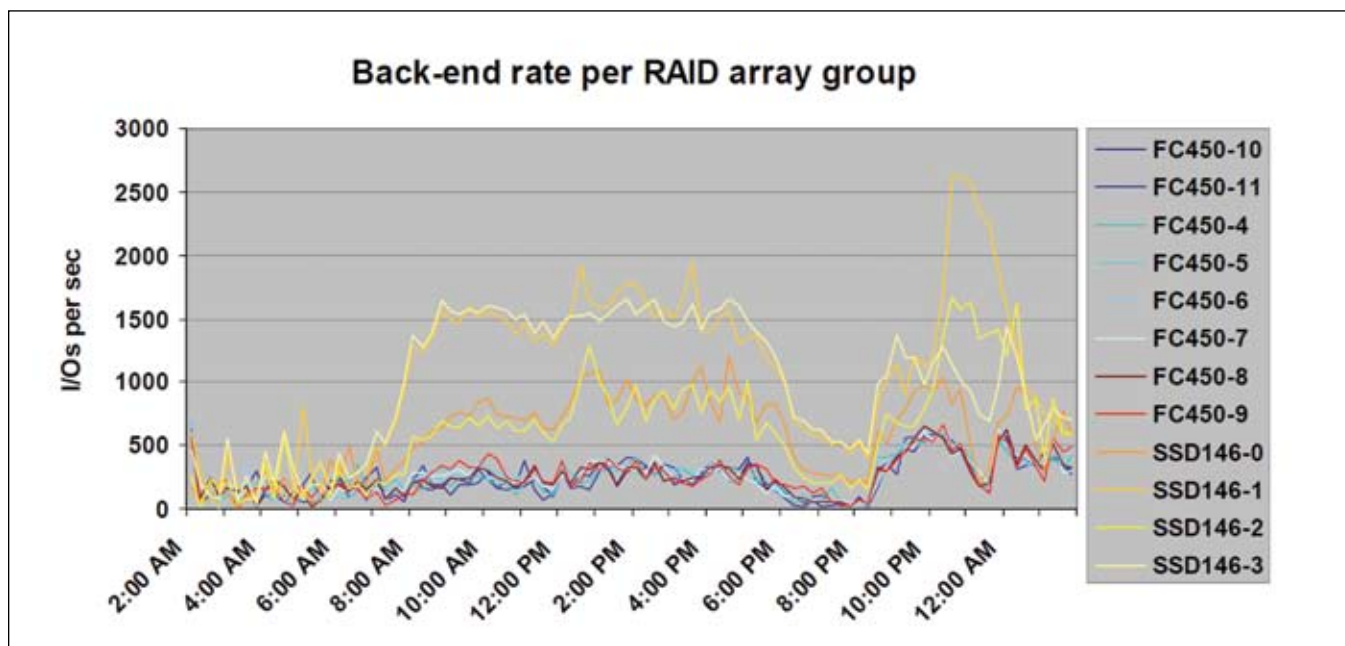


Figure 3: A Balanced Two-Tier Solution With SSDs and Fibre Drives

disk to be serviced, so your measured disconnect time will be 20 times as low as the measured disk response time. A disk response time of 20 ms would show up as a contribution to disconnect time of only 1 ms.

**4. List capacities and capabilities:** To fully understand the space requirement, you need to know the usable space provided by each disk and RAID type combination. This is easily computed by knowing the disk capacities and how many parity drives, and possibly spares, the different RAID schemes use. The properties of the disk and RAID schemes can be entered into a spreadsheet to calculate the effective capacities.

For the performance capabilities, you need to know the response time of each disk and RAID type configuration as a function of the I/O rate. Figure 2 shows this relationship for 10k RPM FC disks, for 15k RPM FC disks, and for Serial ATA (SATA) drives. Note that the drive size isn't shown; the performance

Pool	Size (TB)	Back-end rate
Work	5	5,000
Prod db	2	4,000
Other db 1	20	6,000
Other db 2	10	3,000
Batch data sets	20	5,000
User data	20	1,000
Archive/inactive	40	400
<b>Total</b>	<b>117</b>	<b>24,400</b>

Figure 4: Workload Breakdown

Drive Type	RAID arrays	HDD busy	Back-end rate	Disk resp. time (ms)	Cost
146GB 15k	117	15%	24,400	7.06	117

Figure 5: Current Single-Tier Solution

Drive Type	RAID arrays	HDD busy	Back-end rate	Disk resp. time (ms)	Cost
300GB 15k	59	30%	24,400	8.6	89
450GB 15k	39	45%	24,400	10.9	78
400GB 10k	82	30%	24,400	12.8	82

Figure 6: Three Single-Tier Alternatives

Drive Type	RAID arrays	HDD busy	Back-end rate	Disk resp. time (ms)	Workloads	Cost
146GB 15k	18	48%	12,000	11.5	W+Pr+db2	18
400GB 10k	42	29%	12,400	12.7	db1+B+U+A	42
<b>Total</b>	<b>60</b>		<b>24,400</b>	<b>12.2</b>		<b>60</b>

Figure 7: Two-Tier Solution

## Example of Multiple Tiers vs. Single Tier

The design criteria in this example are:

- The back-end drives should be 50 percent busy at most.
- The time to serve a read miss should be no more than 13 ms.

The assumed relative cost of the drives is listed in Figure 1. Figure 4 shows the workload used for this example.

### Single-Tier Solutions

The best-performing solution is a single tier of 146GB 15k drives. The number of RAID array groups is constrained by the needed capacity. This is a highly expensive option with a relative cost of 117. For this configuration, we can compute the statistics shown in Figure 5. Less-costly, single-tier solutions can

be created by using 300GB 15k, 450GB 15k, or 400GB 10k drives. The relative costs are lower for all these solutions, and the disk response times are higher, but still within the target response time. The large number of 400GB RAID array groups is driven by the performance criterion for response time. For the required capacity, only 45 RAID array groups is sufficient. Figure 6 shows the statistics for these three single-tier alternatives.

### Two Tiers

With two tiers, you can use a mixture of 146GB 15k and 400GB 10k drives for a cost of 60, as shown in Figure 7. This presents a savings of 23 percent over the single-tier 450GB solution, in spite of a higher drive count. The expected disk response time is slightly higher than this for a 450GB single-tier solution.

—GH & WO

is the same regardless of the size.

At a 15 ms target maximum response time, 15k RPM FC drives can handle almost twice the work that a 10k RPM FC drive can (see Figure 2). SATA response times are high even with a light load. Most mainframe installations should be cautious with SATA, as it

forms an entirely different level of performance.

For each disk technology and RAID scheme combination, you need to list the maximum supported number of back-end I/Os per RAID array group. This will be used to match the target of maximum disk busy. The vendors' maximum I/O rates assume small seeks and may not be usable for these purposes. We estimate a RAID array group of 15k RPM drives can do 1,400 back-end operations at 100 percent busy and a group of 10k RPM drives can do 1,000 back-end operations.

### 5. Match capabilities with requirements:

Use the performance capabilities, combined with the performance criteria and the workload characteristics to find a lower bound for the number of RAID array groups needed for each disk and RAID type option for your two workload parts. Likewise, look at the space provided per RAID array group and compute how many groups are needed for each option to satisfy the space criterion for the workload part. The higher of the two lower bounds found is the number of RAID array groups required. This results in a list of options that satisfy your require-

ments; you can choose the most attractive of these.

**6. Validate by examining placement of volumes:** The options created in the previous steps are based on high-level data without taking into account whether it's even possible to place that data on the array groups in a way that supports performance targets. You can validate this. All the volumes must be mapped to the proposed tiering solution. You can use the volume level statistics gathered in the first step to assess if the solution is feasible, and the performance targets can be used to see whether the performance is acceptable on a volume level.

Manually mapping all volumes to the RAID array groups in a way that optimizes tiering is practically impossible; there are too many potential combinations. However, tools are available to help you quickly generate charts to determine whether performance requirements are met.

Figure 3 provides real measurement data and shows the back-end rates per RAID array group for a two-tier proposal. The proposed tiers are 146GB SSD drives combined with 450GB 15k RPM FC drives. The four yellow lines show the back-end rates for the 146GB

SSD drive groups; the blue and red lines show the back-end rates for the eight 15k RPM FC drive groups. You can see the solution is well-balanced for the FC drive groups; they all have approximately the same back-end rate at each point. Furthermore, the back-end rates for the FC drive groups reach their peak at about 670 back-end I/Os at around 11 p.m., corresponding to a utilization of just under 50 percent. The four SSD groups handle more than two-thirds of the back-end I/Os and represent only one-seventh of the capacity.

Sometimes due to individual volume workload intensities, it isn't possible to map the volumes so the proposed solution works. In that case, you can increase the number of RAID array groups of the proposed design and try again, or you can pick one of the other options from step 5.

If the final solution, containing the tiered design and the volume mapping to the tiered design, isn't acceptable, you must adjust the assumptions. Either adjust the performance targets or select a different workload split, and then return to step 2 or 3.

### Conclusion

Tiered storage can be less expensive

and more efficient. However, designing tiered solutions is more challenging than designing single-tier solutions and carries additional risks. With the right tools and methodology, you can be confident the tiered design will be robust and perform well. Such tools also help to assess the cost savings and potential of a single-tier vs. multi-tier solution, and create a detailed plan for mapping the volumes to the new storage system. **Z**

### About the Authors

**DR. GILBERT HOUTEKAMER** is an owner and managing director of IntelliMagic. He holds a Ph.D. from the Delft University of Technology, The Netherlands. He has more than 20 years of experience in I/O performance analysis, and has written numerous publications on this topic, including the book *MVS I/O Subsystems*, which he co-authored with Pat Artis. Email: gilbert.houtekamer@intellimagic.net



**WIM OUDSHOORN** is the software architect for the Balance product at IntelliMagic. He earned his master's degree in Mathematics at the University of Amsterdam. Before joining IntelliMagic, he worked at a large ERP vendor, developing optimization tools for production planning. Email: wim.oudshoorn@intellimagic.net

