Disk Systems

and the

Internal Bandwidth Wars

Author:
David J. Sacks

A **Technology Insights** White Paper from IBM

Disk system vendors sometimes draw attention to a disk system's "internal bandwidth". Is internal bandwidth a useful indicator of disk system performance that applications can attain? Is it a useful measure to compare the performance of two different disk systems? This paper discusses answers to these questions.

Notices

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR INFRINGEMENT. IBM shall have no responsibility to update this information.

IBM products are warranted according to the terms and conditions of the agreements (*e.g.*, IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Trademarks

Please send any feedback on this publication to djsacks@us.ibm.com

Published March 18, 2003

Introduction - the allure of using single numbers to compare complex systems

Trying to represent the quality of complex technology by a single number can be very tempting. That would allow the quality of two different systems or components to be compared by simply comparing two numbers. Higher means better/faster/cheaper, lower means worse/slower/more expensive.

This kind of simplified approach has sometimes been applied to computer disk storage systems where many decision makers (and understandably so) would rather not struggle with the complexities of comparing competing products. The problem is that the appeal of simplicity can be a trap leading to erroneous conclusions.

In disk systems, especially high-end disk systems, an attribute called *internal bandwidth* is often cited as a single number that indicates system performance. In practice, however, internal bandwidth can be a very misleading indicator of disk system performance. A good analogy is how the performance of different servers or microprocessors is sometimes compared using their MIPS (Millions of Instructions per Second) specification; because it is not a very useful value for such a comparison, MIPS has become widely known as a Misleading Indicator of Processor Speed[1].

The inadequacy of internal bandwidth as a useful indicator of disk system performance can be explained in multiple ways. Actually, it can be an interesting exercise to identify as many ways as you can why this is so. This will be done here without reference to any specific vendor or disk system product. It is the principles that matter. We'll begin with some definitions and move on from there.

What is bandwidth?

Bandwidth means the maximum possible or theoretical **throughput** supported by a system or isolated component. It is a measure of **how many** units of something can potentially be processed in a given amount of time. It is often expressed in megabytes or gigabytes per second, but could also be expressed as transactions per hour, I/O requests per second, or in other terms.

Is bandwidth the performance a user or application sees?

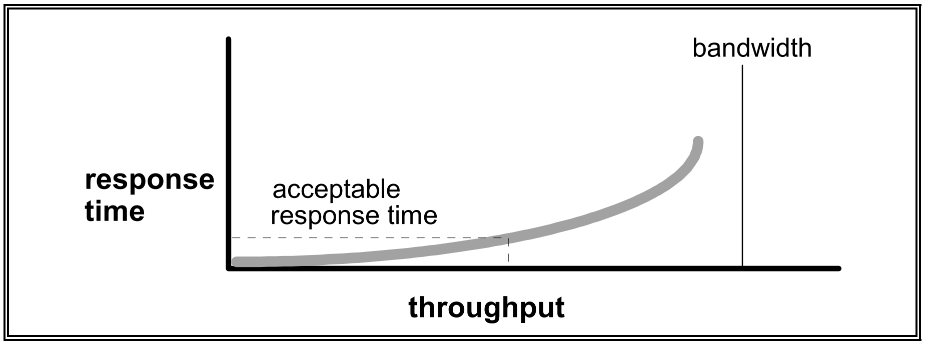
The short answer is no, for several reasons:

- Bandwidth is a number that often appears on component specification sheets, but is rarely if ever an indication of throughput users could experience. For example, a Fibre Channel port on a disk system or host server might have a bandwidth specification of 2Gbit/s, but it will have a lower *sustained or average throughput* in practice due to factors such as protocol overheads, component design and implementation, and workload fluctuations.

[1] Reasons for this include different instruction sets, 32-bit vs. 64-bit architectures, and more.

- Bandwidth says nothing about **response time**. Response time, also called *elapsed time*, measures **how much time** it takes for *a* unit of work, such as an I/O request or a transaction, to complete. Briefly, **bandwidth (or throughput) is how many, while response time is how long**. Moreover, response times computer users (or applications) experience are made up of multiple elements including computer instruction times + I/O times + network times + times spent waiting for these resources. Bandwidth, particularly the bandwidth of an isolated component, does not provide information about *system* response time.

- A system's or component's bandwidth is a fixed value independent of actual workloads; at best it represents a theoretical upper limit to throughput the system or component could handle. In contrast, response time varies with workload throughput; typically, response time of any one unit of work increases as the number of units of work a system or component actually handles increases. Acceptable or desired response time may (and in practice almost certainly does) occur at throughput levels well below the bandwidth value. The following diagram illustrates these relationships; the curve represents the workload input to the component or system (e.g., I/O requests per second).

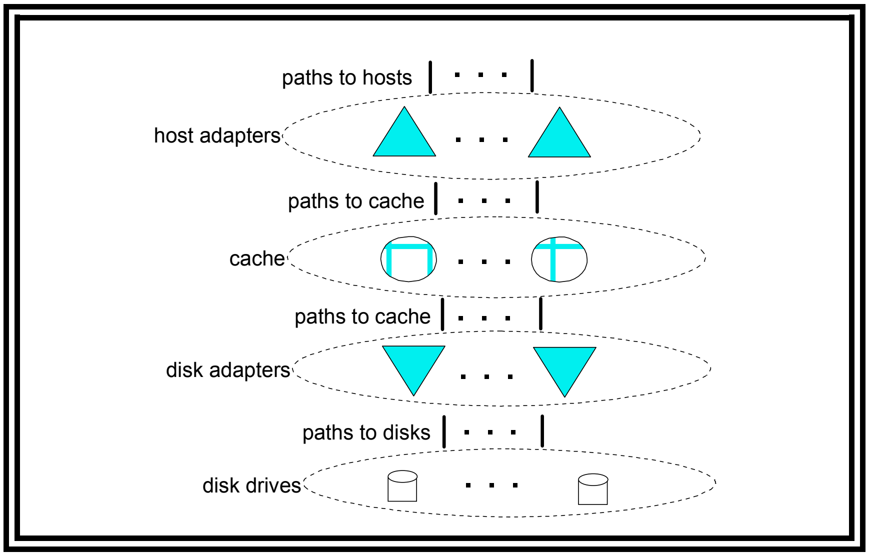


An important consideration is that higher bandwidth (as well as higher throughput in general) does not necessarily imply faster (i.e., lower) response time. For example, a disk system with two disk drives can potentially support twice the bandwidth (i.e., handle twice the requests per second) of another system with only one disk drive; yet, for a single I/O request in isolation, the time to process the request will be about the same in either system (because the single request is handled by a single disk drive and any other disk drives would not be accessed in this case).

In this context, it is useful to note that the term *speed* is ambiguous when referring to computer systems. It could refer to bandwidth (i.e., throughput) or response time or both, depending on the context. When a vendor claims their new system is "faster" or "twice the speed" of other systems, that claim could be about higher throughput, or lower response time, or both.

A basic model of a disk system

As this discussion continues, it will be helpful to refer to a model identifying major elements of a disk storage system - elements through which data moves (as opposed to other elements such as power supplies). In the model, these elements are organized into eight vertical layers: four layers of electronic components shown inside the dotted ovals and four layers of paths (i.e., wires) connecting adjacent layers of components to each other. This model is general enough to apply to most if not all cached disk systems in the marketplace.



Conceptual model of a cached disk system

Starting at the top in this model, there are some number of host computers (not shown) that connect over some number of paths to some number of host adapters. The host adapters connect over some number of paths to some number of cache components. The cache components in turn connect over some number of paths to some number of disk adapters that in turn connect over some number of paths to some number of disk drives. Different disk systems in the marketplace use various terms for the components shown in the model. The number and capacity of these elements, as well as their underlying technology, can vary by system family, product model, and installed configuration.

Here is how a read I/O request is handled in this model. A host issues a read I/O request that is sent over a path (such as a Fibre Channel or FICON path) to the disk system. The request is received by a disk system host adapter. The host adapter checks whether the requested data is already in cache in which case it is immediately sent back to the host. If the data is not in cache, the request is forwarded to a disk adapter that reads the data from the appropriate disk and copies the data into cache. The host adapter sends the data from cache to the requesting host.

What is disk system *internal bandwidth*?

To begin, there is no universally accepted definition of what *disk system internal bandwidth* means, other than a *specification* of the throughput capability of something *inside* a disk system. Internal bandwidth can be contrasted to external performance (i.e., throughput and/or response time) that applications could attain as if the disk system was a "black box".

Referring to the model, it is apparent that application data (and internal control information) flows through multiple layers of components. *Each layer, and each component in a layer, has its own bandwidth.* Some useful points:

1. Any one I/O request generally uses only one component within a layer and may not involve all layers. (Example: a read I/O request passes through one host adapter, and if that request results in a cache hit it does not pass through any of the layers below cache.)

2. The bandwidth of each component within the same layer may or may not be the same. (Examples: In real systems, all disk adapters usually have the same bandwidth. In contrast, some disk systems support multiple types of host adapters (such as 1Gb/s Fibre Channel, 40MB/s UltraSCSI, and 2Gb/s FICON) that can coexist in the same disk system.)

3. The bandwidth of a layer is the aggregate sum of the bandwidths of the components that comprise the layer. (Example: if the layer consisting of the paths to disks has ten paths at 100MB/s each, the bandwidth of that layer is 1GB/s.)

4. The bandwidths of different layers are often different. (Such differences may be appropriate because, due to the nature of applications and disk system designs, some layers will normally have more work to do than others. On the other hand, such differences could indicate potential performance bottlenecks.)

Which particular internal bandwidth is likely to be cited by a vendor?

When disk system vendors cite an internal bandwidth number, they are usually referring to the bandwidth of a particular layer or layers, whether explicitly identified or not. Not surprisingly, the focus is often on the layer with the highest bandwidth. Moreover, when a disk system is offered as a family of multiple models each with various configuration options, the internal bandwidth cited may apply only to the maximum configuration of the high-end model, which may be very different from what a given customer installs.

Drawing attention to the layer with the largest bandwidth may be intended to impress a (less sophisticated) audience, but does not by itself provide much if any information about the external performance of the disk system.

In practice, the highest internal bandwidth generally belongs to either the cache layer or to the two paths-to-cache layers adjacent to cache. This is because most high-end disk system designs require that all data being read from or written to the system must flow through the system's cache.

The lack of clarity in the meaning of disk system internal bandwidth is not only a source of confusion, but a source of potential misunderstanding of product capabilities. Consider an example. Assume the cache layer can deliver up to x GB/s of throughput. If the aggregate bandwidth of the two layers of paths to cache is less than that, then that aggregate path bandwidth is the more meaningful indicator of internal throughput. On the other hand, if the aggregate bandwidth of the two path layers is higher than the bandwidth of the cache layer, then the cache layer bandwidth is the more meaningful indicator. To illustrate, if a hypothetical cache can support up to 1GB/s throughput and is connected to host and disk adapters via ten 1GB/s paths, then a claim of 10GB/s of internal bandwidth would obviously be of little practical significance. The point is that the lower of the two bandwidths, cache or paths-to-cache, would be the more accurate indicator of system internal bandwidth - at least where internal cache activity is concerned.

One storage vendor announced a new disk system with a high internal bandwidth. The details revealed that this was the aggregate bandwidth of the two paths-to-cache layers. The bandwidth of the cache layer itself was actually only 1/4 as large. And both bandwidth numbers applied only to a high-end model of a family of systems.

The lesson is: When a vendor cites a number as "internal bandwidth", ask what layer or layers are being referred to. Even if a vendor explicitly uses the term *cache bandwidth*, it is prudent to ask: Is that the bandwidth of the cache layer or of the two paths-to-cache layers? Whatever the answer, ask what is the bandwidth of the other element because the smaller value is the more meaningful one. And, even with this understanding, do not draw conclusions about external system performance.

How might system design impact internal bandwidth?

Disk system designers can select among various technologies for communicating among different components and layers. Options include buses, switches, direct paths, and combinations of these. A given vendor may claim the approach in the product they are selling is superior to competitors' approaches. Nevertheless, from a customer's perspective, even if the implementation consists of rabbits, carrots, and conveyor belts (with at least two of each for redundancy), as long as the system meets requirements for external performance and system reliability, such "under the covers" design differences may be of little practical importance to anyone other than engineers.

A system's design, and resulting performance capability, is not only about which hardware technology is employed, but how that hardware is used or managed by internal system software. Comparing different systems' internal bandwidth specifications alone, even assuming they apply to comparable internal components or layers, says nothing about how the bandwidths are used. Yet that consideration have can a large impact on performance and on determining whether comparing two systems' internal bandwidth values is even reasonable. Some examples:

1. Some disk systems use cache and paths-to-cache to support RAID-5 parity updates that consist of multiple internal data transfers, while other disk systems offload that activity to the disk adapters. A disk system using cache and paths-to-cache for parity updates may need higher bandwidth for those layers of components just to provide the performance another system achieves through a different design.

2. Some internal volume replication facilities use a "resynchronization" algorithm to periodically move all accumulated changes from a source volume to an associated target

volume to bring the target up-to-date. This traffic may move over internal paths-to-cache. A typical once-a-day resynchronization schedule moves about twenty-four hours worth of changes to each target volume. In contrast, some disk system volume replication facilities use an alternate approach based on a "copy-on-first-write" algorithm that can dramatically lower the traffic that must flow between the source and target, reducing paths-to-cache utilization.

3. Every I/O request has some amount of internal "overhead" processing associated with it. The impact of this overhead will be lowest for sequential workloads that consist of a relatively small number of I/O requests for large blocks of data. But the impact of this overhead will be higher for I/O workloads such as online transaction processing (OLTP) that issue a relatively large number of I/O requests for small blocks. When systems or components differ in overhead efficiency, the difference will be more evident as the number of I/O requests increases. However, bandwidths are often specified as megabytes or gigabytes per second as if the system or component was processing one large continuous unit of work with no overhead at all. The point is, a system or component with a higher bandwidth but a higher overhead per request may not handle random I/O workloads as efficiently as a system or component with a lower bandwidth and lower overhead per request. These overheads are generally not included in product specifications.

The point is that a higher component or layer bandwidth does not in itself indicate better performance than a lower, corresponding bandwidth in another system. Resource management - how the layers and components are used - may be even more important.

What system attributes besides internal bandwidth can impact performance?

There are many other factors which, alone or in combination, impact disk system external performance:

- Architecture vs. implementation vs. configuration. A disk system's architecture, which may simply be a document, presents one picture of potential performance. For example, an architecture may specify a maximum internal or external bandwidth objective of x GB/s. Some or all models of the product actually being manufactured may implement a bandwidth lower than the architectural maximum. The system configuration of the model being installed by a given customer may deliver lower bandwidth than other system configurations. A product brochure or vendor presentation may or may not clearly distinguish these differences.

- Multiple hardware components in the data path. The work done to satisfy any one application I/O request flows through multiple components. Data may flow through the host adapters, cache, disk adapters, disk drives, and various paths between those components. And different I/O requests may use different component layers. It is the *interaction* of all these layers, much more than the attributes of any one layer in isolation, that impacts system performance.

- Host port aggregate bandwidth. Clearly, the highest possible throughput applications could even theoretically obtain from a given disk system is the aggregate bandwidth of the host ports (on the host adapters) installed on the disk system. For example, if a disk system has four host ports at 2Gbit/s each then the aggregate 8Gbit/s, about 800MB/s, is an absolute upper limit to performance applications could potentially see, regardless of the performance characteristics of the rest of the system. Further, as noted earlier, sustained throughput is almost certainly less than aggregate bandwidth. Reasons for this include protocol overhead, and that some disk systems may use host adapters each with multiple ports where adapter design may not allow all ports on that adapter to be 100% busy at the same time.

- Resource utilization skew. In the real world, some disk system components tend to be busier than others. Except for artificial workloads, it is almost impossible to perfectly balance utilization across all components in a large disk system. Some hosts drive more I/Os than others; some data is accessed more frequently than other data. And these skews can vary by time. If activity to a subset of components ever becomes too high, system performance could become unacceptable even though other resources in the system are not very busy.

  Skew is often found in the disk drive and disk path system components. For example, a large portion of I/O requests could be directed to one physical disk or set of disks containing frequently accessed data, and so implicitly directed to the disk path(s) shared by a set of such disks. A system's internal data layout can potentially reduce the skew of disk drive utilization. If data within logical volumes is striped across multiple physical disks, then the occurrence of disk drive performance "hot spots" is reduced. The inverse is also true: if data isn't striped or isn't efficiently striped, the occurrence of hot spots may increase. Some disk systems stripe all data automatically, while in others it is an option that may require manual planning. Striping efficiency varies with implementation.

- RAID-5 implementation. Potential RAID-5 benefits, compared to disk mirroring (RAID-1), include lower cost and increased capacity scalability. But a conventional RAID-5 design has an associated "write penalty" (of up to 3 extra internal I/Os) for each application write I/O request. A given disk system design may implement various efficiencies that significantly reduce this penalty. As one example, if data is efficiently striped the RAID-5 design can take advantage of that striping to significantly reduce the write penalty for sequential data streams by collecting successive writes in cache and destaging them to disk together as if they were one large write.

- Disk paths and speed. If some percentage of application I/O requests need to be satisfied from disk, as is almost certainly the case, then the speed, number, and protocols of the disks and disk paths will be another performance factor. In particular, higher-capacity disks can help reduce system costs and increase capacity scalability, but they reduce potential disk I/O parallelism compared to a larger number of smaller capacity disks. Each internal path to disks usually supports multiple disks, meaning contention for access to that shared resource likely increases as the I/O workload increases. An FC-AL path, for example, may support dozens of disks, but only one disk can transfer data over the path at one time.

- Cache management. Cache can improve the performance of many I/O requests by replacing the delays caused by electromechanical disks with the faster access times of electronic memory. The higher the cache *hit ratio*, the lower the average response time applications see. Differences in cache management efficiency can impact system performance even for caches of the same nominal size. Some examples:

    - Caches often are divided into fixed-size "slots"; one or more slots are used to hold a contiguous block of data. Caches with relatively small slots generally use cache space more efficiently than caches with relatively large slots. For example, a cache with 4KB size slots will waste no space storing 4KB or 8KB size blocks of data, while a cache with 32KB size slots will waste the majority of its space storing those same size blocks, resulting in a lower overall cache hit ratio and thus slower application performance.

    - Another cache efficiency issue is the amount of data staged into cache due to a read request that results in a cache miss operation. Staging too much data into cache adds to disk, disk path, disk adapter, and path-to-cache utilization without any benefit, and also takes up cache space that would be better used for other data. Staging in too little data could increase the likelihood of a subsequent I/O to read nearby data resulting in a cache miss rather than a hit. Some disk systems dynamically adjust the amount of data staged into cache based on ongoing monitoring of application I/O patterns.

    - Cache may be used for various internal activities that are "behind the scenes", reducing the amount of cache space and bandwidth available to satisfy application requests. For example, some system designs use the cache to pass internal information among system components as if the cache were a mail box.

  The point is that different system designs differ in their cache management efficiencies, with a potentially significant impact on application performance, independent of internal bandwidth considerations.

- System parameters. Disk system parameters that impact ongoing system operation can affect maximum application throughput. For example, most modern-day high-end disk systems have an option to protect new data written to cache against loss due to a cache component failure by keeping a second copy in another cache. Turning that option off, as some vendors sometimes do for performance benchmarks, can improve performance but is not a configuration customers may be willing to accept for production systems.

- Performance accelerators. Some disk systems offer what can be called specialized performance accelerators. For example, some disk systems support the SCSI Command Tag Queuing protocol in their host adapters and/or internal disks. Some disk systems improve IBM mainframe I/O performance through a special facility called Priority I/O Queuing. These kinds of capabilities can improve performance and help administrators better manage application service level agreements.

With so many factors that can impact system performance, and with different systems having different designs that are affected by these factors in different ways, it is a major challenge, to say the least, to try to assess external system performance based on considering or comparing internal design elements or bandwidths in isolation.

Internal bandwidth claims can far exceed potential external performance

Internal bandwidth numbers are generally many times higher than throughput applications attain in practice. In one case, a vendor's Web site reports achieving sustained external throughput for a disk system that is about only 13% as large as the internal bandwidth specification claimed for that same system. In another case, a different vendor's graph reflecting disk system performance for a given workload showed external throughput at well under 10% of claimed internal cache bandwidth and at under 2% of claimed cache path bandwidth.

External performance - claims and benchmarks

External disk system performance - throughput and response times applications and users can potentially attain - is the only performance that really matters to a customer. But accurately estimating external system performance capabilities is not simple, and requires that the buyer beware.

Vendor-produced benchmarks may be helpful - or may be of little if any value. Some vendors report the performance results of benchmarks based on 100% cache hits, often all reads of the smallest allowed data blocks (generally 512-byte blocks in UNIX and Intel-based hosts). Or, vendors may report the results of a sequential benchmark finely tuned to maximize sequential throughput. In contrast, realistic customer workloads usually consist of a mix of reads and writes, cache hits and misses, random and sequential access, and are too dynamic to allow a system to be optimally tuned at all times.

The Storage Performance Council (www.storageperformance.org) is a multi-vendor standards organization that maintains and promotes vendor-neutral disk system benchmarks that have been run by many vendors with results publicly documented at the SPC web site. While not a comprehensive measure of system capabilities, the SPC-1 benchmark can help customers better compare the performance characteristics of different disk systems. Customers must judge for themselves the credibility of a vendor's claim of superior performance when that vendor refuses to publish SPC-1 results.

What factors, besides performance, could be included in a disk system decision?

Focusing for a time on one system design topic, as this paper has done, can sometimes make one forget about the bigger picture. Performance is only one of many values a disk system can provide. Assuming two different disk systems meet or exceed a customer's performance requirements, it is helpful to remind ourselves of many other factors that can be important in a disk system evaluation.

- Scalability (maximum usable RAID-protected data, maximum host ports, maximum hosts, etc.)

- Availability (hardware fault tolerance, online changes, online upgrades, concurrent repair, etc.)
- Functions (such as internal copy and remote copy features)
- Management capabilities
- Preference for a single vendor or multiple vendor environment
- Which disk systems are already installed and the customer's experience with those systems
- Total Cost of Ownership
- Vendor experience in the target environment (midrange, mainframe)
- Vendor's overall storage strategy
- Vendor services offerings
- Quality/reputation of the vendor's customer product service/maintenance organization
- The ability of the vendor to provide "one-stop shopping" and/or a systems perspective for IT solutions that can include hardware, software, management, and services.

Summary

Disk systems are complex technology. Measuring disk system performance can be a complex task, whether evaluating one system in isolation or comparing different systems. No single number can completely characterize the performance of a disk system. That is particularly true, as we've seen, for attributes of components in isolation such as internal bandwidth.

Major points discussed in this paper include:

- Disk system internal bandwidth does not indicate or predict external (application) bandwidth
- Disk systems have multiple layers of components with different internal bandwidths; vendors often draw attention to only the highest internal bandwidth
- For a given product/product family, internal bandwidth can vary by system architecture, model, and specific configuration
- Many other elements other than internal bandwidth significantly impact disk system performance
- Vendor-neutral benchmarks, such as SPC-1, can provide an indication of system performance that applications could attain on different disk systems

The value of disk system internal bandwidth in a product comparison or buying decision appears to be low at best.