Jon's Performance Musings: Musings

Jon E. Schmidt
Transaction Design, Inc.
San Rafael, CA, 94901, USA
1.415.256.8369
inform@banbottlenecks.com

Jon is the founder of Transaction Design, Inc. (TDI), a consulting firm located in the San Francisco Area which specializes in capacity/performance studies with clients worldwide. He is the creator of the Ban Bottlenecks® service and has an extensive background in the implementation, testing, and tuning of high-availability systems.

"Colloquium"

I've been developing a two-day "colloquium" (a conference where experts give papers and discuss a topic) on performance for an East-coast bank. Frankly, it's impossible to distill 30 years of experience into two days. On the other hand, this has forced me to organize my thoughts and put down on paper the methodology that I use, and the principal touch-points when I approach performance reviews of a system.

I've also unexpectedly enjoyed doing the background research for the paper. While the technology has boosted the "speeds and feeds" tremendously over the years, the principles of performance and capacity analysis remain the same.

The Jevons Paradox: Technological progress that increases the efficiency with which a resource is used tends to increase (rather than decrease) the rate of consumption of that resource.

Disks Are Bottlenecks

My experience has been that disks are usually a bottleneck. We see it at clients all the time. We work with them to balance their drives and manage disk cache intelligently. I will make two strong statements:

- The best disk environment for OLTP is a set of directly-attached drives operating in a RAID 0 + 1 (mirrored and striped) mode.
- Never use a shared Storage Area Network (SAN) for OLTP transaction files (and certainly not Network Attached Storage).

NonStop OS does an excellent job with its disks. Mirroring, no problem. Striping, yes, we would like to see it. And by striping I'm referring to having a logical disk span multiple physical disks such that a single file on the logical volume gets spread across the physical disks on a block-by-block basis. That way high-volume processing on a single file gets spread across multiple sets of heads.

The other thing we'd like to see is the NonStop disk process (for each disk) able to be replicated across cores. Don't think that's happened yet.

SAN: Strategic And Necessary

Let's not kid ourselves. The world we live in mandates that enterprises go to SANs. The economics are too compelling. The regulatory environment demands archiving of massive amounts of data. The availability and disaster recovery requirements mandate multiple sites with fault- and disaster-tolerant data stores. Management of a SAN can be simpler than management of a data pool per application or per image. RAID, snapshots, and DR synchronization without the involvement of the host are major advantages.

SANs and NASs Are Networks

On the other hand we have to keep in mind that the new storage paradigm is a network. It's a bit-serial network which means that only one packet or segment or transmit unit can be on the wire at a time. So then, yes, you can have multiple wires. But inevitably then you have switches and routers. And with multiple wires, indeed a "fabric", you'll have hops, queues, collisions or drops, and retransmits.

SANs By The Numbers

SAN networking is fast. Very fast. We're talking 1 to 10 Gigabits per second on a wire and probably soon to be faster. With multiple wires the aggregate throughput is much higher. But consider:

- A single modern disk drive is capable of up to 200MB/second sustained transfer rate. Its local cache (in the drive electronics) can transfer data faster, from 300-600MB/second depending on the interface. That means a single disk theoretically can saturate a 2Gb fibre channel by itself.
- SANs are intended to service multiple computer systems (hosts). Therefore the disk access, SAN controller, cache memory, and network fabric are shared.
- A single set of physical drives may in fact contain several logical drives, with each logical drives servicing a different host.
- Controller cache management and drive access scheduling within the SAN may be optimized for throughput, not the random OLTP access that our world needs. It makes no sense for a drive or SAN controller to read a megabyte or so into drive or SAN cache if all the application requires is a single random block of data.
- Therefore our poor little (but extremely critical) OLTP system may be competing against several big batch hosts that may be "sucking up" resources. One of our vendors reports times of up to 10 seconds for a request to come back from the SAN.

Dedicated SANs are Good

Many of these issues go away with a SAN dedicated to the OLTP environment. Competition certainly does. But the bandwidth issues remain and must be considered carefully. On the other hand the throughput advantages of data striping and non-hosted backups, snapshots, and DR are important. Note too that these have to be carefully managed since they will have a performance impact on OLTP access.

You will also want to review the caching policy of the SAN in light of the fact that the NonStop OS knows what it's doing with its disk cache.

Watch For Competition

If you are running SAN disks you need to be proactive and watch for slowdowns. The Measure counters for disks include reads, writes, request-qtime, read-busy-time, write-busy-time, and device-qbusy-time, among others.

I suggest you periodically snapshot a busy or critical time of day. Establish the ratio between the request counts and the qtimes that are associated with those counts. If there is competition within the SAN causing a slowdown, the qtimes will increase even though the request counts did not. We've seen it. Also do the math: Number of I/Os times the blocksize will tell you the fabric bandwidth consumed.

SAN Sanity

SANs can be a tremendous advantage when done right. Retain your sanity and use them intelligently.