RMG Consultants, Inc.

Papers on Library Automation

Performance of Turnkey Library Systems: A Contractual Approach

by Rob McGee

RMG/CAVAL
123 Dover Street,
Richmond, Victoria, AUS 3121
03-427-1288
Fax #: 03-428-5429

RMG Central (Corporate)
333 W. North Ave., Suite F
Chicago, IL 60610
312/321-0432
Fax #: 312/321-9594

RMG East
P.O. Box 1130
Bethesda, MD 20817
301/469-5900
Fax #: 301/469-0823

Performance Warranties

Most of the turnkey automated library systems being sold today may be characterized as online, transaction-processing, database-oriented systems. Within the library automation industry the term "system performance" usually refers to the response-times with which systems perform interactive transactions, although other performance factors should be considered as well. In assisting libraries to avoid performance problems with new or upgraded systems, the approach of RMG Consultants, Inc. (RMG) is as follows: to develop through a deliberate planning and procurement process the best possible understandings between buyer and seller of both the library's requirements and the vendor's system, and to address through explicit contractual warranties the following aspects of System performance that can contribute decisively to the success or failure of a library automation project:

- "Response Times," for both peak and non-peak transaction volumes;
- "Transaction Mixes," representing distributions of transaction types (e.g., charges, Boolean searches, bibliographic record inputs) that the library is likely to experience through successive phases of system implementation;
- "Transaction Throughout Capacity," projecting the maximum volume of transactions that the system can perform during a specified period;
- "Storage Capacity," specifying the capacity of data storage devices, expressed in bytes, or numbers of library records of various types;
- "Database Load Rates," specifying the system's rates for batch-loading bibliographic records and/or authority records in both dedicated and non-dedicated batch modes;
- "Functionality," specifying the functional capabilities of the system — i.e., the applications logic for input, processing, and output of data;
- "System Availability," specifying the "uptime" during which the system will be operable during periods of scheduled use;
- "System Growth Path," specifying how and at what costs a system may be expanded to provide additional capacity throughout its projected implementation phases and beyond.

Performance Factors

In order to guarantee system performance, the buyer and seller should agree upon definitions of performance levels and methods of performance measurement. Because the state of a transaction-oriented database system constantly changes as transactions are performed, files are updated, indexes are expanded, new subsystems are implemented, more terminals are deployed, and transaction volumes increase, the contractual provisions for definition and measurement of system performance should anticipate system growth by allowing for different measures and repeated points of measurement during a system's life cycle. Here the concept of "system state" is particularly useful: not only must phases of system implementation and growth be anticipated, but periods of normal and peak transaction volumes as well. Figure 1 illustrates how transactions might distribute among the subsystems of an integrated library system. This depiction shows how a turnkey system, initially used as a "circulation system," might grow to perform nearly four times as many online transactions with nearly five times as many terminals as other subsystems for acquisitions, cataloging, serials control, and online public access catalog are added. As subsystems are implemented in phases, the number and type of transactions — the "transaction mixes" — may change from week-to-week, month-to-month, and year-to-year until implementation is completed, which for an integrated system can take two to three years, or longer.

Apart from the changes in system workload that accompany implementation phases, systems will experience daily, hourly, and minute-by-minute fluctuations in volumes and mixes of online transactions. Transaction processing makes split-second changes in performance demands upon system resources whose activities are usually measured in milliseconds or microseconds: disk accesses that read or write data; the transfer of bits of information between disk and main memory, main memory and CPU, and CPU and input/output data channels; the amount of main memory used by a program, and the number of CPU instructions it executes; etc. In combination these and other performance factors determine the length of time required for a computer to process a transaction: the interval between its entering and leaving the CPU.

The response-times experienced by users at terminals are even greater because of the time required to transmit data between the CPU and terminals. The transmission rates of telecommunications components

FIGURE 1
Prototype Estimates of Distributions of Online Transactions and Terminals for an Integrated Library System in an Academic Library

(1) Library Subsystem	(2) Number of Online Transactions	(3) Percent	(4) Number of Online Terminals	(5) Percent
Circulation	1.0X	25.6	1.0Y	20.4
Acquisition	.3X	7.7	.3Y	6.1
Cataloging	.3X	7.7	.3Y	6.1
Public Access Catalog	2.0X	51.3	3.0Y	61.2
Serials Control	.3X	7.7	.3Y	6.1
TOTAL	3.9X	100.0	4.9Y	99.9

Note: Comparable estimates for a public library would result in different ratios and distributions of transactions and terminals.

(usually measured in millions, thousands, or hundreds of bits-per-second) will vary from the high-speed transfer between a computer data channel and input/output ports (measured in millions or thousands of bits-per-second) to the baud rate settings of the terminals themselves, ranging in current library systems from 300 baud with slow-speed dial-up devices to 19.2 kilobaud speeds with an emerging class of terminals directly connected to a computer or to a high-speed local area network or communications system. Communications components between an outbound CPU data channel or port and a terminal may include various combinations of front-end communications processors (computers that are designed specially to handle data communications), modems, multiplexors, local area networks, and so forth, each operating at speeds measured from millions to hundreds of bits-per-second.

When poor response-time performance is detected, as much as possible needs to be learned about the system's state at the time, in order for a vendor to identify and correct performance problems. How to identify a period of poor performance and describe the state of system usage at the time is basically a problem of measurement.

Performance Measures

In considering system performance, the buyer and seller must agree upon a definition of "response-time." Typically, in contracts for turnkey library systems, response-time is defined as the amount of elapsed time between the operator at a terminal signaling the computer to perform a transaction (e.g., by depressing an "ENTER" key, wanding a barcode label) and receipt at the terminal of the first character of the system's response. To verify performance using this definition of response-time typically requires someone at each terminal to measure with a stopwatch the interval of time between the operator's input and the system's response. (It may be argued that the last character of the system's response, rather than the first, should be the concluding point of measurement. This gets into issues of system design that are addressed later.)

This type of definition and measurement of response-time grew out of the historic inability of minicomputer operating systems to monitor and measure internal system performance in ways that the more complex operating systems of mainframe computers can do.

Their sophisticated operating system software plus the use of special system-level performance monitoring software allows some mainframe computers to measure system performance of applications (or modules) with respect to transaction volumes and mixes, CPU seconds, main memory utilization, disk inputs/outputs, characters transmitted, lines printed, etc. — and the intervals of time between transactions entering and leaving the CPU. Where such facilities exist, the operation of a library system can be monitored, measured, and reported by the computer system of itself with respect to its own internal performance but not that of the telecommunications system.

Where internal system performance measurement capabilities are available, RMG recommends that they be investigated to determine their acceptability to both buyer and seller as a measurement tool by which the fulfillment or failure to meet contractual response-time performance warranties can be determined. The concepts, provisions, and language of such warranties necessarily depend upon the measurements that are possible.

In the case of turnkey library systems that are implemented with microcomputer and minicomputer operating systems where performance monitoring and measurement tools characteristically are not yet available, or that inadequately address the buyer's interest, other methods of performance audits must be sought. Typically, the established and labor intensive alternative of stopwatch measurements at terminals is used. In RMG's work with clients and vendors, uses of special performance-monitoring devices that attach to data transmission circuits occasionally have been considered, but have not proven practical for monitoring overall system performance at all terminals.

Performance Measurement vs. Benchmark Testing

It seems useful in this discussion to distinguish between measuring the response-time performance of a system as it undergoes normal use, and stress-testing the performance capacity of a system under contrived conditions.

The techniques of measurement that have been considered above — system self-monitoring and measurement and stopwatch measurements of response-time by personnel at terminals — are procedures that can be conducted at any point in a system's life cycle, e.g., at the outset of a new implementation phase or during periods of poor performance. Testing the sufficiency of a system once it is installed in the library is acceptable if the system is known to have adequately met comparable performance levels elsewhere.

However, in situations of risk, a library may not want to enter a contract or pay for a system until it has been proven to perform acceptably under specifically stated conditions of use. The test of a system, in these circumstances is usually referred to as "benchmark testing." A benchmark test is usually performed by stopwatch measurements of operators at terminals who perform different types of transactions according to a script that prescribes transaction volumes and mixes which the system must perform with acceptable response-times. Scripting and supervising a benchmark test to ensure that at each terminal an operator is performing the required transactions and that accurate measurements are being recorded is a matter of considerable logistics,

especially if large numbers of terminals are involved. When such a test is necessary, it is certainly desirable to have the vendor arrange its performance on a system installed elsewhere. Once a system is installed in a library, it will be difficult to remove.

Now that so many turnkey library systems have been installed, it is often possible to obtain proof of system performance at customer installations, or to conduct benchmarks at vendor or customer sites. The absence of performance data from installed customer sites is in itself cause enough to be wary in this marketplace. Buying a system without a relevant performance history is a definite risk.

Although seldom used in the library industry, other techniques for predicting performance include the simulation of a system using computer software models and benchmark testing by executing a stream of prestored transactions with an actual system. In the latter case, copies of real transactions stored on magnetic tape or disk, or perhaps even in microcomputers, can be "played" into the system as a benchmark test.

Benchmark testing, by definition, requires that a system be exercised with existing software. In an industry where most products are still under significant development, it is obvious that the benchmark performance of unavailable software modules cannot be tested. RMG's approach is therefore to develop contracts that anticipate growth in the uses and sizes of systems, with appropriate performance warranties for each phase in a system's life cycle — including implementation of new modules — accompanied by agreed-upon performance measures. If a vendor is unwilling to warrant performance of a system throughout its life-cycle or unable to demonstrate how the transaction throughput capacity of a system may be expanded or its performance enhanced, then let the buyer beware. Performance warranties have little meaning without recourse to a proven data processing growth path for system expansion.

Performance Fixes

RMG's contractual approach to poor system performance is to have the vendor fix it. Usually, a performance problem must be defined rather specifically before it can be corrected. Its occurrence may need to be reproduced so that the conditions of system use that accompany it — i.e., the system state — can be described and the problem observed. Corrections to performance problems may involve hardware or software fixes, or both. Software fixes might vary from "tuning" the allocation of system resources (e.g., assigning more main memory, changing the distribution of data files over disk drives, etc.) to changes in system code, to redesign of a function or module or, in the worst case, the entire system. Hardware solutions might involve expanding the main memory, CPU size, disk storage, etc.

In cases of system outages that exceed guaranteed levels of system uptime, refunds of maintenance charges are customary, but consistent hardware failures may also warrant replacement of some or all components.

To minimize disputes, RMG prefers contracts that specify what action the vendor will take to cure performance problems and a time frame for resolution.

Issues of System Design

One might wonder how vendors can develop systems without knowing how performance will turn out or without techniques to predict the effects of certain design choices. Unfortunately, the success of a system design is for many more an art than a science. The proof of a bad design is the redoing of a system. This may explain many of the late product deliveries within our industry.

Until evidence of success is clear, concern for the performance of library systems will surely focus on the development of online catalogs, replete with issues of user-friendliness, authority control, cross-references, response-times for Boolean queries, huge databases, and hundreds of users.

It is worth noting that while a system can be designed to give quick response-times, it might not provide a high transaction throughput capacity. For example, is it better to have a 5-second response-time for a single-term subject search entered as a single-command transaction through a keyboard or to have 2-second response-times each for a series of 5 menu selections with a touch-terminal or keyboarding of a numbered choice?

Particularly for online public access catalogs, the basic designs of library systems will continue to evolve until the "right" balances among user protocols and system performance factors are reached. This may in some products require so much change that later versions may be quite different from earlier ones, to the extent that the original performance warranties may become an issue. Buying a product that evolves into something quite different, and perhaps unwanted, is a phenomenon to anticipate. The impact upon system performance of widespread use of Boolean queries likely will bring these issues to the fore.

Price/Performance

The concept of price-performance — what a system does, and how much it costs to do it — is fundamental to the attitude toward system performance taken in this paper. What is inside the "black box" and how it works internally is not the focus: How much it costs and what it does for the user is the issue. With this view the buyer can approach the measurement and warranty of turnkey system performance with a concern only for system externals, e.g., verifying the functions, measuring response-times, calculating uptime, and fixing the costs of a system at various sizes and levels of performance, etc. RMG recommends that contracts define how and when system performance may be measured, so that buyer and seller agree in advance on the procedures for acceptance tests, for measurements of the system at any time when poor performance is suspected, and on the nature and time frame allowed for remedies.

Insurance Against Poor Performance

The best insurance against poor performance is to be sure that the purchase contract for a system specifies the library's performance requirements and the vendor's warranties to meet them, and that the buyer and seller enter the agreement with their best understandings and judgments. Payment schedules keyed to acceptance of the system in stages upon proven levels of performance are basic. A performance bond that guarantees successful fulfillment of a contract, else compensation to the buyer, is further assurance. If a seller cannot or will not agree to such contractual provisions, a buyer might be well advised to look elsewhere.