SHARE

SHARE SESSION REPORT

61	M505	Establishing a Performance Management Systems	225
SHARE NO.	SESSION NO.	SESSION TITLE	ATTENDANCE
Computer Management and Evaluation		Lois Palucki	NTC
PROJECT		SESSION CHAIRMAN	INST. CODE

Northern Trust, 125 S. Wacker Dr., Chicago, IL 60675 (312) 630-6275

SESSION CHAIRMAN'S COMPANY, ADDRESS, AND PHONE NUMBER

The evaluation of a data center is based on its ability to provide service to its customers. This evaluation requires: (1) that goals and/or service level agreements have been established, (2) that methods are in place to measure performance against the goals, and (3) that a reporting mechanism is put in place to inform all parties of the level of service being provided. This paper describes the Performance Management System Service Level Objectives, and Monthly Performance Management Reporting system developed at Blue Cross Blue Shield of Indiana.

ESTABLISHING A PERFORMANCE MANAGEMENT SYSTEM

James C. Crane
Blue Cross Blue Shield of Indiana
120 W. Market St.
Mailpoint 7E
Indianapolis, Indiana 46204

Installation Code: BCB

Sponsor: Computer Management and Evaluation Project

Session Number: M505

ABSTRACT

The evaluation of a data center is based on its ability to provide service to its customers. This evaluation requires: (1) that goals and/or service level agreements have been established, (2) that methods are in place to measure performance against the goals, and (3) that a reporting mechanism is put in place to inform all parties of the level of service being provided. This paper describes the Performance Management System, Service Level Objectives, and Monthly Performance Management Reporting system developed at Blue Cross Blue Shield of Indiana.

PURPOSE

The purpose of a performance management system is to have a method to measure and evaluate the performance of a computer installation and to report the results to management. This reporting mechanism must provide information on the current status as well as future projections needed to make strategic planning decisions. Very careful planning and evaluation of the performance management process will make this a very effective tool for management.

DEVELOPMENT

The development of the performance management system will involve many different areas within the company. These areas will establish service level agreements for work to be performed in the data center. However, before the service level agreements can be written, the terms in which to specify the objectives need to be established and understood.

The first task is to identify the workloads that will be present and the resources available. Our environment will have TSO, production and test CICS, production and test INTERCOM, and production and test batch. Once all of the measured workloads have been identified, the next task will be to identify the terms in which to specify objectives for each category.

Starting with TSO, a number of terms come to mind. The first would be availability. Availability can be based on a specified number of hours per day or week, may be limited to certain hours of the day for measurement (prime time), could be based on the number of terminals operational or logged on, or simply on whether the TSO task was started. Response time is the next measurement and you must decide whether to base response time on all transactions or just a subset (trivial). If only a subset is chosen, what method will be used to determine the subset? This could be based on resource consumption or simply a percentage of all transactions or both. Another consideration might be the ratio between different types of transactions. The next item to consider measuring would be transaction volumes. These measurements may be by day (total), by transaction type, or by terminal or user. The last item we have identified is the number of concurrent users. We can measure the average number of users or maximum number of users by hour or day.

The next workload to analyze is CICS. There is a production system as well as a test system and each possible objective will have to be evaluated as far as its applicability to production or test. Availability again becomes our first consideration. We now have a different set of qualifications. We can base availability on the basis of files being opened or closed, transactions being started or stopped, terminals being operational, a target number of hours per day or week, or any combination of these. Response time can be identified by transaction, by specific application, by inquiry or update, and the calculation might be based on average response times or percentage completed within a time constraint. Message volume will again need to be measured since it will become an integral part of any service level objectives. A great deal of consideration will need to be given to the contrasts between the production and test requirements in the above areas.

The INTERCOM workload, being similar to CICS, will have many of the same considerations as CICS; however, because the measurement systems are not as precise, the objectives may be more general.

The final workload to consider is batch. Again there will be two categories: production which typically is long running, high output generating jobs critical to the corporate business; and testing which is normally short running development activity. As a result it is difficult to develop a standard set of requirements to satisfy both. Production batch measurements might include percent of jobs which start and/or end within some tolerance of the scheduled time. This requires some sort of scheduling system be maintained. The time a job is on the print queue waiting output may be an indication of output problems or scheduling conflicts. Another method of measuring the effectiveness of the production batch environment might be based on the delivery of output to the end user on time. To some, it is not important when the job runs just so the output is where it belongs on time. Some other items considered might be total number of jobs or job steps, the number or percentage of system or user abends, the number or percentage of JCL errors, or even the total number of printed lines or pages of output. As you can see, there is room for plenty of imagination in the development of methods of determining performance in the area of production batch. Test batch, on the other hand, is probably more directed to fast, multiple turnarounds. One important question will be whether all tests are treated equally or more likely there will be multiple classes. This classification may be based on resources requested, the department who is requesting the work, the individual requesting the work (i.e. the President), or maybe even based on a costing algorithm or system. Should a standard be based on the quantity of jobs completed or the percentage turned around in a selected time limit? Next you might define turnaround time. Will turnaround time include only the execution, or will it also include printing and/or delivery? One major difference between production and test batch workloads is the ability to predict arrival rates or demand. Development people find it very difficult to predict their demand on the system.

Now that we have identified some of the items or terms that performance measurements may be made in for the different workloads, we must do the same for our resources. The resources we may be interested in monitoring are CPU, channels, DASD, mass storage, tape drives, printers, terminals and memory (paging).

The first item we may wish to measure for CPU's is availability. There are many factors that effect CPU availability; some are hardware related while others are software related. If CPU availability is determined to be a measured objective, the installation will have to define the terms in more detail. The next item will be CPU utilization which may be measured for performance purposes or may only be monitored as a barometer of growth against capacity.

The monitoring of channels will be based on items such as service times, activity counts, percent channel busy, and queue lengths. These same indicators are also valid for DASD monitoring. For tape drives the primary interest would be for tape mounts, outages, and jobs delayed waiting on tape drives. The IBM 3850 MSS (Mass Storage System) items to monitor are availability, staging time, number of cylinders staged, and number of cartridge picks. On the question of availability, it must be determined how much of the 3850 needs to be available due to the multiple components. Terminal availability is a fairly straight forward item to measure, but network availability and response times can also be considered. Memory measurements primarily concern themselves with virtual storage constraint and real memory shortages. Real storage measurements can be measured through the paging subsystem. The last of the hardware resources is the printers and again the first item of concern is availability.

Other factors to monitor would be number of lines or pages printed and queue delays which cause backlogs.

It should be obvious that there is no shortage of values that could be measured and reported on. Most companies probably have someone monitoring most of the items listed. You will most likely find that many individuals are involved and there is little communication between them and no common reporting. This is where the formal (or informal) performance management system becomes important. The next section discusses the development of service level agreements to establish the measurements which will become the basis for the performance management reporting system.

OBJECTIVES

We are now ready to establish the standards for service, and there are four primary terms being used in data processing today for this. The terms are: Service Level Mandates, Service Level Contracts, Service Level Agreements, and Service Level Objectives. Let's take a look at each one individually before we begin setting our standards.

Service level mandates are very easy to define. These are skillfully prepared standards established by high levels of management that become the unquestioned primary goals. All other agreements must fit into the remaining resources after these goals have been met.

Service level contracts are the most demanding of the other three, and they can even be legal contracts. In this type of contract, every facet of the agreement must be established including any penalties for abuse or lack of service. Also it is common that when additional service is required, support (financial) comes from the requester. The provider will usually live by the exact "letter of the law" in refusing service beyond that which is contracted. Both parties have something to gain or something to lose.

Service level agreements on the other hand have a much looser definition of service and performance. Again there are two parties involved, and the requester will expect the service level to be maintained; but usually you will find the pressures of missed standards will be dealt with less harshly. The requester has a level of service he would like to attain but will usually negotiate the service objective to a level the provider can supply.

Service level objectives will normally be set by the operations staff to provide a target for consistent service. Any pressures for missing these targets are usually self imposed and may be used as input to the capacity planning function.

Which one of the above methods, or combination of methods, that will be used will be determined by the organization. The functions of hardware and software acquisition will enter into the decision of what method to use. There is normally a correlation between the degree of commitment and the detail of the measurements.

The majority of the standards we have established fall into the category of service level objectives because no limits have been established for demand from the requester. The two notable exceptions are TSO response time and test batch turnaround, which could be considered service level agreements; however, this is a liberal interpretation. Let's look at each objective individually and determine what its benefits might be.

1. Provide 98% CPU availability. This measures primarily the dependability of the hardware and, secondarily, the software and operational standards. Attainment leads to the smooth running of the computer installation.

2. Provide 98% mass storage availability. This is a measurement of the hardware performance and has the greatest impact on the TSO and test batch workloads.

3. Provide 96% availability for production T.P. This measures the stability of our online systems. A missed standard here can create a work backlog in the user areas and can severely impact the performance of the corporation. There are actually two standards, one for INTERCOM and one for CICS.

4. Provide 95% of all T.P. transactions in 4 seconds or less. This measures the consistency of our online systems. A missed standard here creates a continuity problem in the user areas and can severely impact the day-to-day business of the corporation. There are actually two standards, one for INTERCOM and one for CICS.

5. Provide TSO availability at 95% during prime time. This measurement tests the stability of the TSO system, which is very heavily depended on for the day-to-day work of both development and support personnel.

6. On-time delivery of reports 95% of the time. This is the measure of service to the client areas. Their specific needs were found to be timely availability of their output.

7. Provide 96% printer availability composite of all IBM 3800 printers. This is primarily a hardware measurement, but it can be used to analyze hardware needs.

8. Meet data entry delivery standard 96% of the time. Since the data entry function can impact the production batch environment it is measured individually. We can track problems to personnel, equipment, or clients.

9. Meet microphotography delivery standards 96% of the time. This measurement is the parallel to the on-time delivery of reports for microphotography.

10. Meet data entry performance standard of $1.68 per 1000 keystrokes. This measurement is important to maintain cost control over the data entry function.

11. Meet microphotography performance standard at a combined cost for source and COM of $.017 per frame. Again this is a cost control standard and needs to be maintained to justify the function in-house.

12. Control TP incident days to 12. An incident day is defined as any day in which more than one outage in duration of 5 minutes or more is experienced. The measurement indicates the disruption of service to the client, loss of time by staff, and is measured separately for CICS and INTERCOMM.

13. Execute 90% of all batch tests in 2.5 hours up to a maximum of 750 per day. This measurement is used to determine turnaround service for the development staff.

14. Provide TSO response time of 88% of all transactions in 2 seconds or less. This standard measures the responsiveness of the TSO system and should be a barometer of productivity in the development area.

15. Control unscheduled IPL's to 175 per year. This measures interruptions in productive work and is examined in conjunction with all other availability standards.

16. Maintain terminal availability at 98% during prime time. The TP system availability must be supported with good terminal availability to be valuable. This measurement examines terminal hardware, network, and lines.

17. Maintain 90% availability on the test TP systems. This is a measurement of the stability of the test TP systems for the development area and is measured separately for CICS and INTERCOMM.

REPORTING

Performance reporting can be thought of as a hierarchial or pyramidal structure beginning at the top of the Corporations and working it's way down through the organization. At the top there are very few measurements that are important. As one moves down through the organization, the measurements become more numerous and more detailed. Each lower level of reporting should support the level above it within the pyramid and where possible, draw a relationship between itself and other areas on the same level. In order for any measurements to be meaningful, the relationship between a lower level and its corresponding higher levels must be known in order to be an effective management tool.

Performance reporting related to computer performance management follows this structure. There are a number of areas monitoring specific performance indicators (MVS, CICS, INTERCOM, hardware, data entry, etc.), and each area has detailed performance reporting. All of these must feed the performance management system to create a central reporting system to represent the division. The following describes a computer performance management reporting system which has its basis at a detailed level, but supports division management by reducing the amount of detail transmitted.

A report is to be produced monthly providing ISD management with information pertaining to the performance of the hardware and software as they relate to established service level objectives and to provide a vehicle for conveying the performance improvement plan to staff members responsible for components that effect performance. The report consists of the following five sections:

1. PERFORMANCE SUMMARY - A chart describing all measured operations performance objectives and the actual performance attained during the current month as well as year to date. This section allows the reader to see all the objectives and results in a concise summary.

2. MISSED OBJECTIVES - This section identifies any missed performance objectives. Each missed objective is then analysed to determine the reason for not meeting the objective and then, if possible, suggestions are made which may prevent the objectives from being missed in the future.

3. POTENTIAL PROBLEMS -The purpose of this section is to alert all parties involved in the performance of the data center to items which may cause future performance problems. This section is very valuable because it points out trends in degrading performance, problems developing in critical systems, shortages in hardware or software resources, or simply to communicate concerns of future bottlenecks.

4. ACTIONS TAKEN - A list of actions taken from a hardware or software standpoint which may have effected performance since the last report as well as the results of those changes. The changes may have been a reactive result of previous performance problems or scheduled enhancements.

5. ACTIONS PLANNED -This section is composed of future items planned for the improvement of performance as well as possible implementation dates. Other topics for this section are suggestions which could improve performance, but need to be studied and evaluated as to value and cost.

SUMMARY

Once the service level objectives have been established, the measurements are in place, and the reporting system is presenting management with data with which to measure the data center's performance, management of the data center becomes much easier. The performance data can be used to: (1) identify bottleneck or problem areas, (2) areas where superb or efficient service is being maintained, or (3) plan for future expansion of the data center. It is important to remember that no planning or tuning effort can be successful without a reporting system for the current environment.

MONTHLY PERFORMANCE MANAGEMENT REPORT

MAY, 1983

Approved By:______________________

Approved By:______________________

Approved By:______________________

This report is produced monthly to provide ISD management with information pertaining to the performance of the hardware and software as they relate to established service level objectives and to provide a vehicle for conveying the performance improvement plan to staff members responsible for components that effect performance. The report consists of the following sections:

PERFORMANCE SUMMARY

A chart describing all measured operations performance objectives and the actual performance attained this month and year to date.

MISSED OBJECTIVES

Identification of missed performance objectives, reason for miss, and suggested solutions for correction.

POTENTIAL PROBLEMS

List of potential or identified bottlenecks and suggested solutions to elimination of same.

ACTIONS TAKEN

List of actions taken since the last report to correct missed objectives or bottlenecks and the results of those actions.

ACTIONS PLANNED

List of future plans for the improvement of performance.

PERFORMANCE SUMMARY

	PERFORMANCE STANDARDS	ACTUAL FOR MONTH	YTD AVERAGE/COMPLETE
1.	PROVIDE CPU AVAILABILITY 98% COMPOSITE OF BOTH PROCESSORS.	99%	99%
2.	PROVIDE 98% MASS STORAGE AVAILABILITY.	99%	99%
3.	PROVIDE TP SERVICE AT 96% EFFECTIVE AVAILABILITY (INTERCOMM).	99%	99%
4.	PROVIDE TP SERVICE AT 96% EFFECTIVE AVAILABILITY (CICS).	98%	99%
5.	PROVIDE 95% OF ALL TP TRANSACTIONS IN 4 SECONDS OR LESS IN A LOCAL ENVIRONMENT (CICS).	96%	97%
6.	PROCESS 95% OF ALL TP TRANSACTIONS IN 4 SECONDS OR LESS IN A LOCAL ENVIRONMENT (INTERCOMM).	99%	99%
7.	PROVIDE TSO AVAILABILITY AT 95% DURING PRIME TIME.	97%	98%
8.	ON-TIME DELIVERY OF REPORTS 95% OF THE TIME.	95%	97%
9 .	PROVIDE 96% PRINTER AVAILABILITY COMPOSITE OF ALL PRINTERS. (3800'S).	99%	98%
10.	MEET DATA ENTRY DELIVERY STANDARD 96% OF THE TIME.	99%	99%
11.	MEET MICROPHOTOGRAPHY DELIVERY STANDARD 96% OF THE TIME.	** 93% **	97%
12.	MEET DATA ENTRY PERFORMANCE STANDARD OF $1.68 PER 1K. KEYSTROKES. THIS ITEM IS REPORTED ONE MONTH IN ARREARS DUE TO BUDGET REPORT LAG.	$1.64	$1.17
13.	MEET MICROPHOTOGRAPHY PERFORMANCE STANDARD AT A COMBINED COST FOR SOURCE & COM OF $.017 PER FRAME. THIS ITEM IS REPORTED ONE MONTH IN ARREARS DUE TO BUDGET REPORT LAG.	$.010	$.011
14.	CONTROL TP INCIDENT DAYS TO 12 FOR INTERCOMM. AN INCIDENT DAY IS DEFINED AS ANY DAY IN WHICH MORE THAN ONE OUTAGE IN DURATION OF 5 MINUTES OR MORE IS EXPERIENCED.	1	2
15.	CONTROL TP INCIDENT DAYS TO 12 FOR C.I.C.S. AN INCIDENT DAY IS DEFINED AS ANY DAY IN WHICH MORE THAN ONE OUTAGE IN DURATION OF 5 MINUTES OR MORE IS EXPERIENCED.	2	5
16.	EXECUTE 90% OF ALL BATCH TESTS IN 2.5 HOURS UP TO A MAXIMUM OF 750 PER DAY.	99%	99%
17.	PROVIDE TSO RESPONSE TIME OF 88% IN 2 SECONDS OR LESS.	92%	91%
18.	CONTROL UNSCHEDULED SYSTEM IPL'S TO 175 PER YEAR (14.5 PER MONTH).	11	69
19.	MAINTAIN TERMINAL AVAILABILITY AT 98% DURING PRIME TIME.	99%	99%
20.	MAINTAIN 90% AVAILABILITY ON THE TEST C.I.C.S. SYSTEM.	** 86% **	90%
21.	MAINTAIN 90% AVAILABILITY ON THE TEST INTERCOMM SYSTEM.	** 84% **	93%

** Indicates a missed standard

MISSED OBJECTIVES

1. Standard 11 missed. MEET MICROPHOTOGRAPHY DELIVERY STANDARD 96% OF THE TIME.
    - Only 93% of the delivery standard was reached in May due to a supply of bad COM film. There were a total of nine cases of bad film and the result was late deliveries due to several reruns.
2. Standard 20 missed. MAINTAIN 90% AVAILABILITY ON THE TEST C.I.C.S. SYSTEM.
    - Test CICS availability was missed due to a communication problem. The availability was calculated using improper time periods.
3. Standard 21 missed. MAINTAIN 90% AVAILABILITY ON THE TEST INTERCOMM SYSTEM.
    - The availability of test ICOM was missed last month because it is not always run if it is not needed for testing and it conflicts with other work that needs to be run. Test ICOM usage has been consistently decreasing since the conversion to CICS began and this standard should probably be dropped.

POTENTIAL PROBLEMS

The major concern of operations today is whether we have the resources to continue to increase the CICS workload, through conversions and new applications, and still maintain the performance standards we have set in the service level agreements. A special group, The CICS Tuning and Performance Committee, has been formed to monitor, tune, and report to management the changes that will be necessary to achieve this task.

The increased utilization of the B083 has shown us that our PAGE/SWAP subsystem is inadequate and we will most likely see a degradation in TSO performance before the installation of the 3380's in July.

ACTIONS TAKEN

The following actions were taken in May to enhance the performance of the data center.

- CICS was split into two separate processing regions operating under the control of the CICS Multiple Region Option (MRO). This provided the primary region with some virtual memory relief which was used for performance improvements as well as capacity for growth of new applications.

- A change was made to the Installation Performance Specifications (IPS) on the A168 to establish new storage isolation values for the online systems. This change was then monitored during the month to determine the final values to be used.
- The VSAM recoverable catalog was converted to the new ICF structure. The results were more efficient processing, performance improvements, and a savings in common virtual storage requirements.
- Much attention has been given to subsecond response time at recent SHARE and GUIDE conferences. A study was conducted to determine the lost hours (dollars) that could be attributed to poor response time (TSO and batch) in the development area. The report shows the possible increase in productivity possible if the necessary resources were available to support the development effort. The report is currently awaiting management approval.

ACTIONS PLANNED

The PAGE/SWAP subsystem on the B083 will be upgraded in July and an upgrade for the A168 needs to be evaluated. Several minor tuning changes were made in May to handle the demand on the PAGE/SWAP subsystem of both processors and keep response times within standard.

A dramatic change needs to be made to the testing system to gain control over the workload and be able to offer adequate service. There has been very little activity on this subject in the last two months, but more effort is planned in the future. Due to other commitments, this may not be done until the next cycle.

Increase the maximum number of address spaces and TSO users. We have added many more terminals to the TSO network without raising the maximum number of users that can be logged on concurrently. The consideration is whether to allow more concurrency at the cost of degraded performance. After installing the new PAGE/SWAP subsystem on the B083, we should be able to make this change without the performance degradation.

As a result of the impending crises with TSO response time before the July implementation of the new PAGE/SWAP devices, an effort needs to be undertaken to try to tune the B083 system around the I/O imbalances. The time and personnel to do this may be difficult to find because of other workplan commitments.

Through routine analysis of the disk subsystem, we have identified several imbalance situations that could be potential performance bottlenecks. During the preparation of the next DASD space plan, several recommendations will be made concerning the use of the IBM 3350 disk subsystem to alleviate these bottlenecks.