UNIT 63 - BENCHMARKING

UNIT 63 - BENCHMARKING

This unit is contains far more information than can possibly be covered in a single lecture.

The middle sections, D and E, contain a detailed technical review of a benchmark model. Depending on the abilities and interests of your students you may wish to omit these sections and move on to the description of the AGT benchmark in section G, or focus the lecture on the technical aspects and omit the descriptive example.

UNIT 63 - BENCHMARKING

A. INTRODUCTION

- benchmarking is a key element in minimizing the risk in system selection
    - often the customer does not have precise plans and needs - these will be determined to some extent by what the GIS industry currently has to offer
        - no vendor's product yet meets the requirements of an ideal GIS
    - customer needs reassurance - real, live demonstration - that the system can deliver the vendor's claims under real conditions
        - the GIS industry is still young, there are two few success stories out there

- a benchmark allows the vendor's proposed system to be evaluated in a controlled environment
    - customer supplies data sets and a series of tests to be carried out by the vendor and observed by the customer
    - an evaluation team is assembled and visits each vendor, performing the same series of tests on each system
    - tests examine specific capabilities, as well as general responsiveness and user-friendliness

- reinforces the written response from the vendor by actual demonstration of capabilities
    - demonstration is conducted in an environment over which the customer has some control - not completely at the vendor's mercy as e.g. at trade show demonstrations

- equipment is provided by the vendor, data and processes must be defined by the customer

- a benchmark can be a major cost to a vendor - up to $50,000 for an elaborate benchmark
    - in some cases part of these costs may be met by the customer through a direct cash payment

Two types of benchmarking

- qualitative benchmark asks:
    - are functions actually present?
    - do they live up to expectations?
    - are they easy to use?

- quantitative benchmark asks:
    - does the proposed configuration have the necessary capacity to handle the planned workload?

Benchmark script

handout - Benchmark script example (2 pages)

- benchmark uses a script which
  - details tests for all of the functions required
  - permits both:
    - subjective evaluation by an observer (qualitative)
    - objective evaluation of performance (quantitative)
- must allow all of the required functionality to be examined
- failure of one test must not prevent remainder of test from being carried out
- must be modular
  - customer must be able to separate the results of each test
- conditions must be realistic
  - real data sets, realistic data volumes

B. QUALITATIVE BENCHMARKS

- in the qualitative part of the benchmark it is necessary to evaluate the way the program handles operations
- functions cannot be evaluated simply as present or absent

overhead - Qualitative assessment

- functions are not all equally necessary - they may be:
  - necessary before any products can be generated, e.g. digitizing
  - necessary to some products but not others, e.g. buffer zone generation
  - necessary only to low-priority products, i.e. nice to have

C. QUANTITATIVE BENCHMARKS

- in quantitative tests, procedures on problems of known size are executed
  - analysis of results then establishes equations which can be used to predict performance on planned workload
  - e.g. if it takes the vendor 1 hour to digitize 60 polygons during the benchmark, how many digitizers will be needed to digitize the planned 1.5 million polygons to be put into the system in year 1?
  - this is known in computer science as performance evaluation

Performance evaluation (PE)

- developed in the early days of computing because of need to allocate scarce computing resources carefully

a subfield of computer science

- requires that tasks be broken down into subtasks for which performance is predictable
- early PE concentrated on the machine instruction as the subtask
  - specific mixes of machine instructions were defined for benchmarking general-purpose mainframes
    - e.g. the "Gibson mix" - a standard mixture of instructions for a general computing environment, e.g. a university mainframe
- multi-tasking systems are much more difficult to predict because of interaction between jobs
  - time taken to do my job depends on how many other users are on the system
  - it may be easier to predict a level of subtask higher than the individual machine instruction
  - modern operating systems must be "tuned" to perform optimally for different environments
    - e.g. use of memory caching, drivers for input and output systems

Subtasks for GIS PE

- specifying data structures would bias the benchmark toward certain vendors
  - e.g. cannot specify whether raster or vector is to be used, must leave the choice to the vendor
- similarly, cannot specify programming language, algorithms or data structures
- a GIS benchmark must use a higher level of subtask
  - an appropriate level of subtask for a GIS benchmark is:
    - understandable without technical knowledge
    - makes no technical specifications
  - e.g. "overlay" is acceptable as long as the vendor is free to choose a raster or vector approach
  - e.g. "data input" is acceptable, specifying digitizing or scanning is not
- therefore, a GIS PE can be based on an FRS and its product descriptions, which may have been generated by resource managers with no technical knowledge of GIS

Requirements for a quantitative benchmark

- need a mathematical model which will predict resource utilization (CPU time, staff time, plotter time, storage volume) from quantities which can be forecast with reasonable accuracy
  - numbers of objects - lines, polygons - are relatively easy to forecast
  - technical quantities - numbers of bytes, curviness of lines - are less easy to forecast
- the mathematical form of the model will be chosen based on expectations about how the system operates

- e.g. staff time in digitizing a map is expected to depend strongly on the number of objects to be digitized, only weakly on the size of the map (unless large maps always have more objects)
- requires a proper balance between quantitative statistical analysis and knowledge about how the procedures operate

GIS PE is more difficult

- GIS PE is more difficult than other types of PE because:
  - uncertainties over the approach to be adopted by the vendor (data structure, algorithms)
  - high level at which tasks must be specified
  - difficulty of forecasting workload
  - no chance of high accuracy in predictions
    - however even limited accuracy is sufficient to justify investment in benchmark

D. EXAMPLE MODEL OF RESOURCE UTILIZATION

- this section describes a mathematical model developed for a quantitative benchmark

overhead - Model of resource utilization

handout - A model of resource utilization

Subtasks

- begin with a library of subtasks L
  - this is the set of all GIS functions defined conceptually
  - e.g. overlay, buffer zone generation, measure area of polygons, digitize

Products and data input

- FRS identified a series of products
  - identified as R1, R2,...,Ri,...
- each product requires a sequence of subtasks to be executed
- data input also requires the execution of a series of subtasks for each dataset, e.g. digitize, polygonize, label

Frequency required

- each product is required a known number of times per year
  - Yij is the number of times product i is required in year j
- knowledge extends only to the end of the planning horizon, perhaps year 5

Execution of tasks

- execution of a subtask uses resources
  - e.g. CPU, staff or plotter time

- these can be quantitatively measured
  - e.g. CPU time measured in seconds
  - e.g. staff time in minutes
  - note: indications are (Goodchild and Rizzo, 1987) that staff time (human) is more predictable than CPU

    time (machine) because of complications of computer accounting systems, multitasking etc.

- Mak is the measure of resource k used by subtask a
  - k is one of the resources used
  - a is one of the subtasks in the library L

Prediction

- in order to predict the amount of resources needed to create a product, need to find a mathematical relationship between the amount of resource that will be needed and measurable indicators of task size
  - e.g. number of polygons, queries, raster cells, lines

- Pakn is predictor n for measure k, subtask a

- Mak = f(Pak1,Pak2,...,Pakn,...)
  - e.g. the amount of staff time (Mk) used in digitizing (a) is a function of the number of polygons to be digitized (Pak1) and the number of points to be digitized (Pak2)

- the general form of the prediction function f will be chosen based on expert insight into the nature of the process or statistical procedures such as regression analysis
  - e.g. use the results of the benchmark to provide "points on the curve" with which to determine the precise form of f

Forecast

- given a prediction function, we can then forecast resource use during production with useful, though not perfect, accuracy

  Wkit is the use of resource k by the tth subtask required for a single generation of product i

  Wki = sum of Wkit for all t is the amount of the resource k used by all subtasks in making product i once

  Vkj = sum of (Wki Yij) for all i is the amount of resource k used to make the required

numbers of all products in year j

Summary of phases of analysis

overhead - Summary of phases of analysis

1. Define the products and subtasks required to make them

2. Evaluate each subtask from the results of the qualitative benchmark

3. Analyze the system's ability to make the products from the qualitative evaluations in (2) above

4. Obtain performance measures for known workloads from the results of the quantitative benchmark

5. Build suitable models of performance from the data in (4) above

6. Determine future workloads

7. Predict future resource utilization from future workloads and performance models, and compare to resources available, e.g. how does CPU utilization compare to time available?

E. APPLICATION OF MODEL

- this section describes the application of this model of resource use in a benchmark conducted for a government forest management agency with responsibilities for managing many millions of acres/hectares of forest land
- FRS was produced using the "fully internalized" methodology described in Unit 61
- FRS identified 33 products
  - 50 different GIS functions required to make them out of a total library of 75
- GIS acquisition anticipated to exceed $2 million

Three phases of benchmark

1. data input - includes digitizing plus some conversion of existing digital files

2. specific tests of functions, observed by benchmark team

3. generation of 4 selected products from FRS

- these three phases provided at least one test of every required function
- for functions which are heavy users of resources, many tests were conducted under different workloads

- e.g. 12 different tests of digitizing ranging from less than 10 to over 700 polygons

Qualitative benchmark

- each function was scored subjectively on a 10-point scale ranging from 0 = "very fast, elegant, user-friendly, best in the industry" to 9 = "impossible to implement without major system modification"
  - score provides a subjective measure of the degree to which the function inhibits generation of a product
  - maximum score obtained in the set of all subtasks of a product is a measure of the difficulty of making the product

Quantitative benchmark

- since this was an extensive study, consider for example the quantitative analysis for a single function - digitizing
- digitizing is a heavy user of staff time in many systems
- delays in digitizing will prevent system reaching operational status
  - digitizing of complete database must be phased carefully over 5 year planning horizon to allow limited production as early as possible
- as stated above, benchmark included 12 different digitizing tasks
- resource measure of digitizing is staff time in minutes
- predictors are number of polygons and number of line arcs
  - line arcs are topological arcs (edges, 1-cells) not connected into polygons, e.g. streams, roads
  - other predictors might be more successful - e.g. number of polygons does not distinguish between straight and wiggly lines though the latter are more time-consuming to digitize - however predictors must be readily accessible and easy to forecast
- sample of results of quantitative benchmark: polygons line arcs staff time (mins) 766 0 930 129 0 136 0 95 120
- benchmark digitizing was done by vendor's staff - well- trained in use of software, so speeds are likely optimistic

Model

overhead - Models of time resources required

- expect time to be proportional to both predictors, but constants may be different

m = k1p1 + k2p2 m is measure of resource used p is a predictor - p1 is polygons, p2 is line arcs k1, k2 are constants to be determined

Results

- the equation which fits the data best (least squares) is:

  m = 1.21 p1 + 0.97 p2

  - i.e. it took 1.21 minutes to digitize the average polygon, 0.97 minutes to digitize the average line arc

- to predict CPU use in seconds for the digitizing operation:

  m = 2.36 p1 + 2.63 p2

  - i.e. it took 2.36 CPU seconds to process the average polygon

- uncertainties in the prediction were calculated to be 34% for staff time, 44% for CPU time
  - suggests that humans are more predictable than machines

- adding together staff time required to digitize the forecasted workload led to the following totals: Year Time required (minutes) 1 185,962 2 302,859 3 472,035 4 567,823 5 571,880 6 760,395

- the average working year has about 120,000 productive minutes in the daytime shift
  - by year 6 the system will require more than 6 digitizing stations, or 3 stations working 2 shifts each, or 2 stations working 3 shifts each
  - this was significantly higher than the vendor's own estimate of the number of digitizing stations

    required, despite the bias in using the vendor's own staff in the digitizing benchmark

F. LIMITATIONS

- difficult to predict computer performance even under ideal circumstances
  - GIS workload forecasting is more difficult because of the need to specify workload at a high level of generalization
  - the predictors available, e.g. polygon counts, are crude

- the model is best for comparing system performance against the vendor's own claims, as implied by the configuration developed in response to the RFP
  - it is less appropriate for comparing one system to another

- it assumes that the production configuration will be the one used in the benchmark
  - staff will have equal levels of training
  - hardware and software will be identical
  - it is difficult to generalize from one configuration to another - e.g. claims that one CPU is "twice as powerful" as another do not work out in practice

- however, any prediction, even with high levels of uncertainty, is better than none
    - after a quantitative benchmark the analyst probably has better knowledge of system performance than the vendor

G. AGT BENCHMARK EXAMPLE

Project Background

- in 1983, Alberta Government Telephones (AGT) had been operating a mechanized drawing system for 5 years
    - however, lack of "intelligence" in automated mapping system was increasingly hard to justify given growing capabilities of GIS
    - management was showing interest in updating the record-keeping system

- an FRS and RFP for an AM/FM system were developed by a consultant in cooperation with staff
    - three companies were identified as potential suppliers and a benchmark test was designed

- tests included placement and modification of "plant" (facilities), mapping, report generation, engineering calculations, work order generation

- tests were designed to be progressively more difficult
    - all vendors were not expected to complete all tests

- data and functional requirements analysis were sent in advance to all vendors for examination
    - actual benchmark script and evaluation criteria were not sent in advance
    - vendors were asked to load supplied data in advance of benchmark
    - methods chosen to load and structure data were part of the evaluation
    - visits were made to each vendor 5 weeks before the actual benchmark to clarify any issues

- providing the data before the script is typical of benchmarks for systems that are primarily query oriented
    - prevents planning for the queries that are presented in the script
    - on the other hand, benchmarks for systems that are product oriented will normally provide the script in advance

- in the AGT case, actual benchmarks were conducted by a team of 3, spending one full working week at each vendor
    - during the benchmark the vendor's staff were responsible for interacting with the system, typing commands, etc.
    - the benchmark team acted as observers and timekeepers, and issued verbal instructions as appropriate

- must recognize that the vendor's staff are more familiar with the system than the typical employee will be during production
    - thus the benchmark is biased in favor of the vendor in its evaluation of user

interaction - the vendor's staff are presumed to be better than average digitizer operators etc.

- during the benchmark, the intent of each phase of testing was explained to the vendor
  - positive and negative evaluations were communicated immediately to the vendor
  - the project team met each evening to compare notes

- a wrapup session at the end of the benchmark identified major difficulties to the vendor, who was invited to respond

- when the three benchmarks were completed the results were assessed and evaluated and became part of the final decision-making stages

REFERENCES

Goodchild, M.F., 1987. "Application of a GIS benchmarking and workload estimation model," Papers and Proceedings of Applied Geography Conferences 10:1-6.

Goodchild, M.F. and B.R. Rizzo, 1987. "Performance evaluation and workload estimation for geographic information systems," International Journal of Geographical Information Systems 1:67-76. Also appears in D.F. Marble, Editor, Proceedings of the Second International Symposium on Spatial Data Handling, Seattle, 497-509 (1986).

Marble, D.F. and L. Sen, 1986. "The development of standardized benchmarks for spatial database systems," in D.F. Marble, Editor, Proceedings of the Second International Symposium on Spatial Data Handling, Seattle, 488-496.

EXAM AND DISCUSSION QUESTIONS

1. Discuss the Marble and Sen paper listed in the references, and the differences between its approach and that presented in this unit.

2. How would you try to predict CPU utilization in the polygon overlay operation? What predictors would be suitable? How well would you expect them to perform based on your knowledge of algorithms for polygon overlay?

3. Since a computer is a mechanical device, it should be perfectly predictable. Why, then, is it so difficult to forecast the resources used by a GIS task?

4. Compare the approach to GIS applications benchmarking described in this unit with a standard description of computer performance evaluation, for example D. Ferrari, 1978, Computer Systems Performance Evaluation. Prentice Hall, Englewood Cliffs, NJ.

5. In some parts of the computing industry, the need for benchmarks has been avoided through the development of standardized tests. For example such tests are used to compare the speed and throughput rates of numerically intensive supercomputers, and of general-purpose mainframes. Are such tests possible or appropriate in the GIS industry?

6. GIS product definition exercise - 2 following pages.

Last Updated: August 30, 1997.