CSC470 Fall, 2000

Computer System Modelling And Analysis

Laboratory #5
November 14th/16th & 21st/23rd, 2000

Purpose

This laboratory involves the performance analysis of software systems. Given a functional description of a software architecture, you are asked to use a combination of analysis and simulation to complete a performance analysis of various alternative system architectures. The system consists of client nodes and server nodes interconnected by a LAN. Each node is composed of a processor and one or more disks.

Context

Consider a computer system to support the marketing operations of a company involved in retail sales. Customers call marketing personnel to make enquiries, place orders, and return goods. The marketing personnel use nodes of the distributed system that can be thought of as *client* nodes. Other functions, which include credit checks, inventory, warehouse, and accounting, are supported on other nodes that can be thought of as *server* nodes. A *local area network* interconnects all the client and server nodes.

Many design decisions are required to specify a complete system design. First, the number of client and server nodes must be chosen, along with the number of disks to be placed at each node of each type. Second, the appropriate multithreading level must be chosen for the server nodes, so that space commitment is minimized yet software servers are prevented from limiting overall system performance.

The characteristics that are known for the various software components are (1) the frequencies with which various software paths are invoked, (2) the average instruction path length for each software component, and (3) the number of I/O operations invoked.

For the hardware components of the nodes, it is necessary to know the usual characteristics: (1) instruction execution rates for processors, (2) the average I/O time at disks, and (3) the average time to transmit a message over the network from one node to another.

Client Software

On behalf of customers, one of three types of processes is invoked on the client servers: Inquiry, Purchase, or Return. Each such process consumes some processor and I/O resources on its own, and it also calls on the required server level processes. The characteristics of client level processes are given in the table below (including their frequencies of calls to software processes on the server nodes).

Type	Trans./hr.	Instr.(M)	# I/O's	V_{crc}	V_{inv}	V_{wrh}	V_{acc}
Inquiry	1000	8	40	--	5	--	--
Purchase	250	40	320	1	3	3	1
Return	50	48	240	--	2	2	1

Server Level Processes

The services that must be provided to support the marketing operation are:

(1). Credit Check
Check the customers credit before allowing a purchase.

(2). Inventory
Check the inventory to determine the current price and the availability of each of the products in a list.

(3). Warehouse
Provide instructions to the warehouse to ship a specific quantity of a specific product to a given customer (or cancel a set of previous ship orders).

(4). Accounting
Indicate that a particular customer is to be billed for an order (or that a prior billing is to be reversed).

The table below gives the resource requirements of each of the server level processes.

Type	Instr.(M)	# I/O's
Credit Check	80	160
Inventory	8	64
Warehouse	48	96
Accounting	64	80

Hardware for Nodes

The server processes are assigned to nodes with relatively powerful processors and fast disks, whereas the client processes generally run on less specialized and powerful (and expensive) equipment. A local area network interconnects all the nodes. Each message that is sent from one node to another must pass over the network, where it may encounter queueing delays.

The important performance parameters of the hardware components to consider initially are given in the table below.

component	service capacity
server processor	40 M instr / sec.
server disk	10 msecs. / I/O
client processor	8 M instr / sec.
client disk	15 msecs. / I/O
network	20 msecs.

Approaches

(1). Since this is an "open" sytem, bounding analysis or utilization analysis will be useful for determining how many nodes of each type will be needed to support the workload.

(2). An analytic model, such as a simple two-level *Method of Layers* model, can be used to estimate the response times to the client requests.
   i. One model can represent each server node and the network, and can be used to determine the time required to complete each server level call;
   ii. A second model can treat the server processes as servers themselves, and can use response time results from the first model to parameterize a *delay* center in the model. Then the model that includes the client processor, the client disk(s), and the delay server can be solved to obtain response time estimates for the client services.

(3). A Csim model can include all the client and server nodes as well as the network to assess the response times.

Performance Targets

Resonse time targets for the client services are:

Inquiries	15 seconds
Purchases	60 seconds
Returns	40 seconds

Assumptions

Here are a number of assumptions to facilitate your modelling task, both with the analytic models and the simulation model.

1. Assume that arriving customer requests are evenly distributed across all the client nodes (i.e., each request is equally likely to arrive at any client node).
2. Assume that client requests to server processes are evenly distributed over all server nodes.
3. Assume that each call from a client node to a server node and each response requires that one message be sent over the network.
4. Client nodes handle requests one at a time in FCFS order (i.e., the next request is not started until the previous request is completed).
5. Server nodes can have multithreading capability so that multiple requests for the same or different services can be processed concurrently. (Multithreading will reduce queueing for the software servers.) Your models should be capable of reflecting the effect of different limits on the degree of multithreading (i.e., how many requests can be concurrently active at a server node).
6. At both client and server nodes, assume that files are placed such that the I/O load is balanced across all disks at the node.

Some Design Questions

Here are some design questions of possible interest:

- Will the system bottleneck be some hardware resource, or the software servers at the server nodes?
- How does the multithreading limit at the server nodes impact the throughput capacity of the system and the response times to client requests?
- How many server nodes are required, and how many disks should each server node have?
- How many client nodes are required, and how many disks should each client node have?
- What will be the response times to the requests made through the client nodes?

Preparation

Use some combination of the approaches described above to answer the kinds of design questions indicated above. Be prepared to apply the same techniques to variations of the situation in which:

i. overall transaction rate is higher or lower,
ii. balance among transaction types is altered,
iii. resource requirements of client or server processes are changed,
iv. hardware component characteristics change, or
v. software is modified, altering its mean instruction path length, or possibly allowing multithreading at the server nodes.