WORKLOAD SELECTION

Gaia Maselli

maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

So far

- State the goals and define the system
- List services and possible outcomes
- Select metrics (procedure)
- Select evaluation techniques
- Select workload

Motivation

- Performance measurements involve monitoring the system while it is being subjected to a particular workload
- In order to perform meaningful measurements, the workload should be carefully selected
- Before performing measurements performance analyst needs to understand:
  - What are the different types of workload?
  - Which workload are commonly used by other analysts?
  - How are the appropriate workload types selected?
  - How is the measured workload data summarized?
  - How can the desired workload be placed on the system in a controlled manner?

The art of workload selection

Terminology

Workload: The requests made by the users of the system (anything that goes in the system)

Workload used in measurements are called ***benchmarks***

Test workload

- Any workload used in performance studies
- Test workload can be *real* or *synthetic*

1. Real workload

- Observed on a system being used for normal operations (packet trace)
- It cannot be repeated ➔ generally not suitable for use as a test workload

Terminology

Test workload

2. Synthetic workload (something made up, much more control over the input)
   - Similar to real workload
   - Can be applied repeatedly in a controlled manner
   - No large real-world data files
   - No sensitive data
   - Easily modified without affecting operation
   - Easily ported to different systems due to its small size
   - May have built-in measurement capabilities.

The art of workload selection

- Workload is the most crucial part of any performance evaluation project
- If the workload is not properly selected ➔it is possible to reach misleading conclusions
- Proper workload selection requires 4 major considerations
  - Service exercised by the workload
  - Level of detail
  - Representativeness
  - timeliness

Serviced exercised (cont)

- The best way to start the workload selection is to view the system a service provider
- Each system provides a number of services, and making the list of services is one of the first steps in systematic performance evaluation study
- System Under Test (SUT) is used to denote the complete set of components
- In every system there is something that you want to change (specific component whose alternatives are being considered Component Under Study - CUS)
- The workload is determined primarily by the SUT (system services determine the workload and the metrics)
- (you do not put the workload directly on the component but on the system)
- The metrics chosen should reflect the performance of services provided at the system level and not at the component level
- The basis for the workload selection is also the system and not the component

Serviced exercised (cont)

- If the system provides multiple services, the workload should exercise as complete a set of services as possible
- In considering the services exercised, take into account the purpose of the study
  - A workload may exercise the most efficient features of the system or the least efficient
  - A workload for email application may bring out the worst in videoconference app
- To summarize: the requests at the service –interface level of the SUT should be used to specify or measure the workload

Services exercised

- N.B. *workload can be **specified** depending on the system level*
- In many systems there is a hierarchy of interfaces at which the requests are serviced
- A single request at a higher level may result in one or more requests at the lower level
- The workload could be described by summarizing the requests at any one of the interface levels, depending upon what constitutes the SUT

Example on a network

- Consider the problem of selecting or designing the workload to compare two networks
- Network layers and corresponding workloads

(A single request at a higher level may result in one or more requests at the lower level)

- *Application* layer consists of applications such as web, mail, FTP, etc.
  - Workload would consist of specifying the frequency of various types of applications and their associated characteristics

- *Transport* layer deals with end-to-end aspects of communication between the source and destination nodes
  - Services provided include segmentation and reassembly of messages, connection initiation, maintenance and disconnection
  - Workload to compare two transport protocols should specify frequency and size of various *messages*, frequency and duration of various type of *connections*

Example on a network

- *Network* layer routes packets from a given source node to a given destination over multiple links
  - Workload to should specify the source-destination matrix, the distance between source and destination, the characteristics of *packet* transmitted
- *Datalink* layer deals with transmission of frames over a single link
  - Workload to compare two datalink protocols should specify lengths and arrival rates of *frames*
- *Physical* layer deals with the transmission of *individual bits* over the physical medium
  - Service: transmission of a *bit* (or *symbol*)
  - Workload to compare two links should match the frequency of various symbols or bit patterns observed on real networks

Level of detail

- After list of services, the next step in workload selection is to choose *the level of detail in reproducing the requests* for these services
- A workload description may be as long as a time-stamped record of all requests or it can be as short as the single most commonly used request
- Possible levels of details:

1. Most frequent requests:
    - select the most frequently requested service and use it as workload
    - Commonly used as initial workload
    - It is particularly valid if one type of service is requested much more often than others or is a major consumer of resources in the system

Level of detail (cont)

2. Frequency of request types
   - To list the various services, their characteristics, and frequency
3. Time-stamped sequence of requests
   - To get a time-stamped record (trace) of requests on a real system and use it as a workload
   - Problem: it may be too detailed, and it is inconvenient for analytical modeling
4. Average resource demand
   - In some cases the resource demands placed by the requests, rather than the requests themselves, are used as the workload
5. Average and/or distribution of resource demands
   - The average demand may not be sufficient in some cases, and it may be necessary to specify the complete probability distribution for resource demands

Level of detail (cont)

- The workload description used for analytical modeling are also referred to as **nonexecutable** workloads since they are not in a form suitable for execution on a real system

- A trace of user commands that can be executed directly on a system is called **executable**

Representativeness

- A test workload should be representative of the real application, i.e., the test workload and the real application match in the following 3 aspects:

1. *Arrival rate*: the arrival rate of requests should be the same or proportional to that of the application
2. *Resource demands*: the total demands on each of the key resources should be the same or proportional to that of the application
3. *Resource usage profile*: the sequence and the amounts in which different resources are used should reflect application needs

Timeliness

- Workload should follow the changes in usage pattern in a timely fashion
- Users behavior has changed considerably over the years
- Users change their usage pattern depending upon the services provided by the new systems
- Workload become obsolete as soon as they become well understood
- It is important that workloads represent the latest usage pattern
- Timeliness is a difficult goal to achieve

Other considerations in workload selection

- *Loading level*: a workload may exercise a system to its
  - *full capacity (**best** case)*
  - *beyond its capacity (**worst** case)*
  - *at the load level observed in real workload (**typical** case)*
- It may be interesting to study the performance of a system under all cases or just one
- Example - Effectiveness of a congestion control scheme, the network should exercise beyond its capacity, while the packet transmission scheme should be tested for normal as well as under heavy load, since retransmissions may be required under both circumstances

Workload characterization

- In order to test multiple alternatives under identical conditions, the workload should be *repeatable*
- Since a real user environment is generally not repeatable, it is necessary to study the real-users environments, observe the key characteristics, and develop a *workload model* that can be used repeatedly

Workload characterization

- Once a workload model is available, the effect of changes in the workload and system can be studied in a *controlled manner* by simply changing the parameters of the model
- There are several techniques for workload characterization

Workload characterization

- User: the entity that makes the service requests at the SUT interface
- **Workload component** or **workload unit** is used instead of the user (e.g., applications, user sessions, station in a network)
- The measured quantities, service requests, or resource demands, which are used to model or characterize the workload, are called **workload parameters** or **workload features**
  - examples: packet sizes, source destinations of packets, number of packets and arrival times

Workload characterization

- There are several method to characterize workload parameters

1. Averaging (presenting a single number that summarizes the parameter values observed)
2. Specifying dispersion (if there is large variability in the data, variance and standard deviation are more appropriate)
3. Single-parameter histograms (a histogram shows the relative frequencies of various values of a paramenter)
4. Markov models

Markov models

- Sometimes it is important to have not only the number of service requests of each type but also their ***order***
- The next request is generally determined by the last few requests
- If it is assumed that the next request depends only on the last request, then the requests follow a ***Markov model***
- Such model can be described by a ***transition matrix***, which gives the probabilities of the next state given the current state
- Transition probabilities give an accurate picture of the order of requests than the frequency of requests

Markov model (cont)

- Example: traffic monitoring on a computre network showed that most of packets were of two sizes – small and large
- Small packets constitutes 80% of traffic. An average of four small packets are followed by an average of one big packet
- Sequence: ssssbssssbssss
- The corresponding transition matrix is:

	Next	packet
Current packet	Small	Large
Small	0.75	0.25
Large	1	0

Homework

- Choose a computer network system for performance study. Describe the system and list:

1. Services and their possible outcomes
2. Performance metrics
3. System parameters
4. Workload parameters
5. Evaluation techniques