Computer Systems: Evaluating Correctness & Performance - II

Presenter: Sandeep K. S. Gupta

Reference:

Arizona State University
School of computing, informatics, and decision systems engineering
Agenda

- System Performance Evaluation – Systematic Approach
- Analytical Performance Modeling
  - Probability & Random Variable
  - Little’s Law
  - M/M/1 Queue
Some Terminology

- **System**: collection of hardware, software and firmware under study
- **Evaluation techniques**
  - measurement, simulation, analytical modeling
- **Metrics**: criteria used to quantify system performance
  - e.g. system throughput, network bandwidth, response time
- **Workloads**
  - scheduler: job mix
  - network: packet mix
  - database: query set
Keep in Mind

- Learn to select appropriate evaluation techniques, performance metrics and workloads for analyzing a system
- Measure and Analyze Appropriately
- Design Effective Experiments
- Validated Simulation
Systematic Approach - I

- State goals and define system
  - define boundaries of system under study: what can be ignored

- List services and outcomes
  - service: network transmits packets from source to destination
  - outcomes: packets delivered correctly, corrupted, lost

- Select metrics for quantitatively evaluating system behavior
  - e.g. speed, accuracy, availability of service

- List system and workload parameters
  - system parameters generally immutable, e.g. CPU speed
  - workload parameter types
    - analytical: probability distribution
    - simulation: traces
Systematic Approach - II

- Select factors to study
  - factors = parameters that vary
  - factor values = levels
  - select a short list likely to be important
  - consider the implications of varying a factor before studying it!

- Select evaluation technique
  - measurement, simulation, analytical modeling
  - appropriate technique depends upon
    - feasibility of measurement
    - time and resources available, accuracy required

- Select workload: list of service requests
  - analytical modeling: probability distributions of requests
  - simulation: web access traces
  - measurement: scripts or program executions
Systematic Approach - III

- Design experiments for effectiveness
  - phase 1: identify important factors (many factors, few levels)
    - fractional factorial experimental designs are effective
  - phase 2: fewer factors, more levels of significant factors

- Analyze and interpret data
  - use statistical techniques to consider impact of variability
  - results and conclusions are separate things!

- Communicate results effectively (high level, no jargon)
  - graphs and charts help distill results but do not stand alone!
### Selecting an Evaluation Technique

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Analytical Modeling</th>
<th>Simulation</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td>any</td>
<td>any</td>
<td>post-prototype</td>
</tr>
<tr>
<td>Time required</td>
<td>small</td>
<td>medium</td>
<td>varies</td>
</tr>
<tr>
<td>Tools</td>
<td>analysts</td>
<td>programs</td>
<td>instrumentation</td>
</tr>
<tr>
<td>Accuracy</td>
<td>low</td>
<td>moderate</td>
<td>varies</td>
</tr>
<tr>
<td>Trade-off evaluation</td>
<td>easy</td>
<td>moderate</td>
<td>difficult</td>
</tr>
<tr>
<td>Cost</td>
<td>low</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Saleability</td>
<td>low</td>
<td>medium</td>
<td>high</td>
</tr>
</tbody>
</table>
Rules of Thumb

- Until validated, all evaluation results are suspect!
  - always validate one analysis modality with another
  - measurements are as susceptible to errors as other techniques
    - experimental error, program bugs
  - beware of counterintuitive results!

- Combining evaluation techniques is useful
  - analytical model: find interesting range of parameters
  - simulation: study performance within parameter range
Commonly Used Metrics

- **Nominal capacity**: maximum achievable under ideal conditions
  - networks: nominal capacity = bandwidth
- **Throughput**: requests / unit time
- **Usable capacity**: max throughput for given response time limit
- **Efficiency**: usable capacity / nominal capacity
- **Utilization**: fraction of time resource busy servicing requests
- **Idle time**
- **Reliability**: probability of error, MTBE
- **Availability**: fraction of time system servicing requests
- **Mean uptime**: MTBF
Types of Performance Metrics

- Higher better (HB), e.g. throughput
- Lower better (LB), e.g. response time
- Nominal is best (NB), e.g. utilization
  - too high: bad response time; too low: resources underused
Setting Performance Requirements

-- (for a design or procurement)

- Requirements should be **SMART**
  - specific (quantitative)
  - measurable (can verify that system meets requirements)
  - acceptable (high enough to be useful)
  - realizable (low enough to be achievable)
  - thorough (requirements should be specified for all outcomes)
Basic Probability concepts
Probability

- **Experiment**: e.g. toss a coin twice
- **Sample space**: set of all possible outcomes of an experiment
  - \( S = \{HH, HT, TH, TT\} \)
- **Event**: a subset of possible outcomes
  - \( A = \{HH\}, B = \{HT, TH\} \)
- **Probability of an event**: a number assigned to an event \( \Pr(A) \)
  - Axiom 1: \( \Pr(A) \geq 0 \)
  - Axiom 2: \( \Pr(S) = 1 \)
  - Axiom 3: For every sequence of pairwise disjoint (mutually exclusive) events

\[
\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)
\]

- Two events **A and B are independent** in case
  - \( \Pr(AB) = \Pr(A)\Pr(B) \)
- If A and B are events with \( \Pr(A) > 0 \), the **conditional probability of B given A** is

\[
\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}
\]
Random variable

- A function defined on a sample space and about whose values a probability statement can be made

Examples of $N(t)$:
- Number of phone calls occurring in an interval of length $t$
- Number of customers waiting in a queue or in service at time $t$
- Number of deaths in an epidemic occurred in $(0, t)$
- Population of a city at time $t$
Random Variable and Distribution

- A *random variable* $X$ is a numerical outcome of a random experiment.

- The *distribution* of a random variable is the collection of possible outcomes along with their probabilities:
  - Discrete case: $\Pr(X = x) = p_\theta(x)$
  - Continuous case: $\Pr(a \leq X \leq b) = \int_a^b p_\theta(x)dx$
Expectation

- A random variable $X \sim \Pr(X=x)$. Then, its expectation is
  
  $$E[X] = \sum_x x \Pr(X = x)$$

- In an empirical sample, $x_1, x_2, \ldots, x_N$,
  
  $$E[X] = \frac{1}{N} \sum_{i=1}^N x_i$$

- Continuous case:
  
  $$E[X] = \int_{-\infty}^{\infty} x p_{\theta}(x) dx$$

- Expectation of sum of random variables
  
  $$E[X_1 + X_2] = E[X_1] + E[X_2]$$
Variance

- The variance of a random variable $X$ is the expectation of $(X - E[X])^2$:

$$Var(X) = E((X - E[X])^2)$$
$$= E(X^2 + E[X]^2 - 2XE[X])$$
$$= E(X^2 - E[X]^2)$$
$$= E[X^2] - E[X]^2$$
Random Variables and Probability Distribution

- **Probability Distribution**
  - For each possible value, \( x_i \), for discrete random variable \( X \), there is a probability of occurrence, \( P(X = x_i) = p(x_i) \)
  - \( p(x_i) \) is the **probability mass function (pmf)** of \( X \), and obeys the following rules:
    1. \( p(x_i) \geq 0 \) for all \( i \)
    2. \( \sum_{all \ i} p(x_i) = 1 \)
Random Variables & Probability Distribution

- The set of pairs \((x_i, p(x_i))\) is the probability distribution of \(X\)
- Example:
  - For a fair die:
    - Probability Distribution:
      - \(\{(1, 1/6), (2, 1/6), (3, 1/6), (4, 1/6), (5, 1/6), (6, 1/6)\}\)
  - For an unfair
    - Probability Distribution:
      - \(\{(1, 1/12), (2, 1/3), (3, 1/6), (4, 1/6), (5, 1/6), (6, 1/12)\}\)
Cumulative Distribution

**Cumulative Distribution Function**

- The pmf gives probabilities for individual values \( x_i \) of random variable \( X \)
- The cumulative distribution function (cdf), \( F(x) \), gives the probability that the value of random variable \( X \) is \( \leq x \), or \( F(x) = P(X \leq x) \)
- For a discrete random variable, this can be calculated simply by addition:
  \[
  F(x) = \sum_{x_i \leq x} p(x_i)
  \]
Poisson distribution

- **Poisson distribution** is a *discrete* probability distribution.
- It expresses the probability of a number of events occurring in a fixed period of time if these events
  - occur with a *known average rate*, and
  - are *independent* of the time since the last event.
Poisson distribution

- The probability that there are exactly \( k \) occurrences (\( k \) being a non-negative integer, \( k = 0, 1, 2, \ldots \)) is

\[
f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},
\]

- \( e = 2.71828\ldots \),

- \( \lambda \) is a positive real number, equal to the expected number of occurrences that occur during the given interval. For instance, if the events occur on average every 4 min, and you are interested in the number of events occurring in a 10 minute interval, you would use as model a Poisson distribution with \( \lambda = 10/4 = 2.5 \).

- Mean: \( \lambda \)

- Variance: \( \lambda \)
(negative) exponential distributions are a class of continuous probability distribution.

used to model the time between independent events that happen at a constant average rate.

probability density function

\[ f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \]

where \( \lambda > 0 \) is a parameter of the distribution, often called the rate parameter. The distribution is supported on the interval \([0, \infty)\).

cumulative distribution function

\[ F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \]

Mean: \( \frac{1}{\lambda} \)

Variance: \( \frac{1}{\lambda^2} \)
In that case where $\lambda$ is taken to be the rate, i.e., the average number of occurrences per unit time, if $N_t$ is the number of occurrences before time $t$ then we have

$$\Pr(N_t = k) = f(k; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},$$

and the waiting time $T$ until the first occurrence is a continuous random variable with an exponential distribution (with parameter $\lambda$). This may be deduced from the fact that

$$\Pr(T > t) = \Pr(N_t = 0) = e^{-\lambda t}.$$
Saying the same thing

- In a Poisson process with rate \( \lambda \), the number of points occurring in a fixed length \( t \) has the Poisson distribution, or equivalently, the lengths of the intervals separating successive points are independent and have identical, exponential distributions.
Poisson dist. example problem

- An interrupt service unit takes \( t_o \) sec to service an interrupt before it can accept a new one. Interrupt arrivals follow a Poisson process with an average of \( \lambda \) interrupts/sec.

- What is the probability that an interrupt is lost?
  - An interrupt is lost if 2 or more interrupts arrive within a period of \( t_o \) sec.

\[
\Pr\{Y \geq 2\} = 1 - \Pr\{Y < 2\} = 1 - P_0(t_o) - P_1(t_o) = \\
1 - \left(\frac{(\lambda t_o)^0 e^{-\lambda t_o}}{0!}\right) - \left(\frac{(\lambda t_o)^1 e^{-\lambda t_o}}{1!}\right) = 1 - (1 + \lambda t_o) e^{-\lambda t_o}
\]
In a bank branch, a rechargeable Bluetooth device sends a “ping” packet to a computer every time a customer enters the door.

Customers arrive with Poisson distribution of \( \lambda \) customers per day.

The Bluetooth device has a battery capacity of \( m \) Joules. Every packet takes \( n \) Joules, therefore the device can send \( u = \frac{m}{n} \) packets before it runs out of battery.

Assuming that the device starts fully charged in the morning, what is the probability that it runs out of energy by the end of the day?

\[
\Pr\{Y \geq u\} = 1 - \Pr\{Y < u\} = 1 - \sum_{k=0}^{u-1} \frac{\lambda^k e^{-\lambda}}{k!}
\]
Poisson Process

- It is an extension to the Poisson distribution in time
- It expresses the number of events in a period of time given the rate of occurrence.
- Formula

\[ P_m(t) = \frac{(\lambda t)^m e^{-\lambda t}}{m!} \]

- A Poisson process with parameter \( \lambda \) has expected value \( \lambda \) and variance \( \lambda \).
- The time intervals between two events of a Poisson process have an exponential distribution (with mean \( 1/\lambda \)).
Number of events between time $a$ and time $b$ is given as $N(b) - N(a)$ and has a Poisson distribution.

Inter-arrival time of event A and event B is given as $\tau_B - \tau_A$ and has an exponential distribution.
How to decide when a system is operating properly?

- Ex: Service Level Agreements (SLA) which guarantees how dependable networking service or power service will be.

Systems alternate between two states of service:

- Service accomplishment (working), where the service is delivered as specified in SLA.
- Service interruption (not working), where the delivered service is different from the SLA.

Failure = transition from state 1 (working) to state 2.

Restoration = transition from state 2 (failed) to state 1.
Quantifying Dependability (2)

- **Module reliability** = measure of continuous service accomplishment (or time to failure).
  - *Mean Time To Failure (MTTF)* measures Reliability
  - *Failures In Time (FIT) = 1/MTTF*, the failure rate
    - Usually reported as failures per billion hours of operation
- **Mean Time To Repair (MTTR)** measures Service Interruption
- **Mean Time Between Failures (MTBF) = MTTF+MTTR**
- **Module availability** measures service as alternate between the two states of accomplishment and interruption (number between 0 and 1, e.g. 0.9)
  - *Module availability = MTTF / (MTTF + MTTR)*
Example: calculating reliability

- If modules have **exponentially distributed lifetimes** (the age of a module does not affect its probability of failure), and that failures are independent, the overall failure rate (FIT) is the sum of failure rates of the modules.

- Calculate FIT (rate) and MTTF (1/rate) for 10 disks (1M hour MTTF per disk), 1 disk controller (0.5M hour MTTF), and 1 power supply (0.2M hour MTTF):

  \[
  \text{Failure Rate} = 10 \times \left( \frac{1}{1,000,000} \right) + \frac{1}{500,000} + \frac{1}{200,000}
  \]

  \[
  = \frac{10 + 2 + 5}{1,000,000}
  \]

  \[
  = \frac{17}{1,000,000}
  \]

  \[
  = 17,000 \text{FIT}
  \]

  \[
  MTTF = \frac{1,000,000,000}{17,000}
  \]

  \[
  \approx 59,000 \text{ hours}
  \]