Workload Characterization Techniques

- Terminology
- Components and Parameter Selection
- Workload Characterization Techniques: Averaging, Single Parameter Histograms, Multi-parameter Histograms, Markov Models, Clustering
- Clustering Method: Minimum Spanning Tree, Nearest Centroid
- Problems with Clustering

Terminology

- User = Entity that makes the service request
- Workload components:
  - Applications
  - Sites
  - User Sessions
- Workload parameters or Workload features: Measured quantities, service requests, or resource demands. For example: transaction types, instructions, packet sizes, source-destinations of a packet, and page reference pattern.

Components and Parameter Selection

- The workload component should be at the SUT interface.
- Each component should represent as homogeneous a group as possible. Combining very different users into a site workload may not be meaningful.
- Domain of the control affects the component: Example: mail system designer are more interested in determining a typical mail session than a typical user session.
- Do not use parameters that depend upon the system, e.g., the elapsed time, CPU time.

Components (Cont)

- Characteristics of service requests:
  - Arrival Time
  - Type of request or the resource demanded
  - Duration of the request
  - Quantity of the resource demanded, for example, pages of memory
- Exclude those parameters that have little impact.

Workload Characterization Techniques

1. Averaging
2. Single-Parameter Histograms
3. Multi-parameter Histograms
4. Markov Models
5. Clustering

Averaging

- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mode (for categorical variables): Most frequent value
- Median: 50-percentile

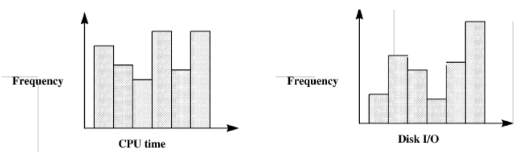
Specifying Dispersion

- Standard deviation *s*: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Coefficient Of Variation: $s/\bar{x}$

s=x

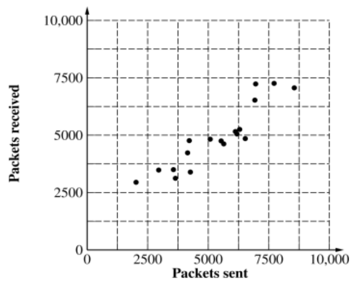
- Minimum and Maximum

Single Parameter Histograms



- *n* buckets £ *m* parameters £ *k* components values.
- Use only if the variance is high.
- Ignores correlation among parameters.

Multi-parameter Histograms

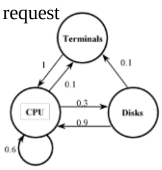


- Difficult to plot joint histograms for more than two parameters.

Markov Models

- Markov
  $\Rightarrow$ the next request depends only on the last request
- Described by a transition matrix:

From/To	CPU	Disk	Terminal
CPU	0.6	0.3	0.1
Disk	0.9	0	0.1
Terminal	1	0	0



- Transition matrices can be used also for application transitions. E.g., P(Link|Compile)
- Used to specify page-reference locality.
  P(Reference module *i* | Referenced module *j*)

Transition Probability

- Given the same relative frequency of requests of different types, it is possible to realize the frequency with several different transition matrices.
- If order is important, measure the transition probabilities directly on the real system.
- Example: Two packet sizes: Small (80%), Large (20%)
  - An average of four small packets are followed by an average of one big packet, e.g., ssssbssssbssss.

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

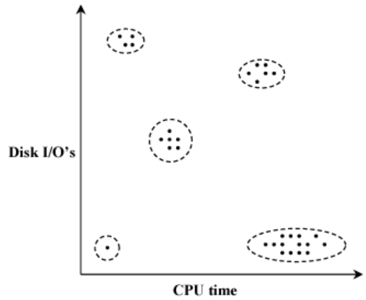
Transition Probability (Cont)

- Eight small packets followed by two big packets.

Current Packet	Next packet Small	Large
Small	0.875	0.125
Large	0.5	0.5

- Generate a random number *x*.
  $x \leq 0.8$) generate a small packet;
  otherwise generate a large packet.

Clustering



Clustering Steps

1. Take a sample, that is, a subset of workload components.
2. Select workload parameters.
3. Select a distance measure.
4. Remove outliers.
5. Scale all observations.
6. Perform clustering.
7. Interpret results.
8. Change parameters, or number of clusters, and repeat steps 3-7.
9. Select representative components from each cluster.

1. Sampling

- In one study, 2% of the population was chosen for analysis; later 99% of the population could be assigned to the clusters obtained.
- Random selection
- Select top consumers of a resource.

2. Parameter Selection

- Criteria:
  - Impact on performance
  - Variance
- Method: Redo clustering with one less parameter
- Principal component analysis: Identify parameters with the highest variance.

3. Transformation

- If the distribution is highly skewed, consider a function of the parameter, e.g., log of CPU time

4. Outliers

- Outliers = data points with extreme parameter values
- Affect normalization
- Can exclude only if that do not consume a significant portion of the system resources. Example, backup.

5. Data Scaling

1. Normalize to Zero Mean and Unit Variance:

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

2. Weights:

$x_{ik}^0 = w_k x_{ik}$

w_k / relative importance or $w_k = 1/s_k$

3. Range Normalization:

$$x'_{ik} = \frac{x_{ik} - x_{min,k}}{x_{max,k} - x_{min,k}}$$

Affected by outliers.

Data Scaling (Cont)

- Percentile Normalization:

$$x'_{ik} = \frac{x_{ik} - x_{2.5,k}}{x_{97.5,k} - x_{2.5,k}}$$

Distance Metric

1. Euclidean Distance: Given $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ and $\{x_{j1}, x_{j2}, \ldots, x_{jn}\}$

$d=\{\sum_{k=1}^{n} (x_{ik}-x_{jk})^2\}^{0.5}$

2. Weighted-Euclidean Distance:

$d=\sum_{k=1}^{n} \{a_k(x_{ik}-x_{jk})^2\}^{0.5}$

Here a_k, k=1,2,…,n are suitably chosen weights for the n parameters.

3. Chi-Square Distance:

$$d = \sum_{k=1}^{n} \left\{ \frac{(x_{ik} - x_{jk})^2}{x_{ik}} \right\}$$

Distance Metric (Cont)

- The Euclidean distance is the most commonly used distance metric.
- The weighted Euclidean is used if the parameters have not been scaled or if the parameters have significantly different levels of importance.
- Use Chi-Square distance only if $x_{.k}$'s are close to each other. Parameters with low values of $x_{.k}$ get higher weights.

Clustering Techniques

- Goal: Partition into groups so the members of a group are as similar as possible and different groups are as dissimilar as possible.
- Statistically, the intragroup variance should be as small as possible, and inter-group variance should be as large as possible.

Total Variance = Intra-group Variance + Inter-group Variance

Clustering Techniques (Cont)

- **Nonhierarchical techniques**: Start with an arbitrary set of k clusters, Move members until the intra-group variance is minimum.
- **Hierarchical Techniques**:
  - Agglomerative: Start with n clusters and merge
  - Divisive: Start with one cluster and divide.
- Two popular techniques:
  - Minimum spanning tree method (agglomerative)
  - Centroid method (Divisive)

Minimum Spanning Tree-Clustering Method

1. Start with $k = n$ clusters.
2. Find the centroid of the i^{th} cluster, $i=1, 2, \ldots, k$.
3. Compute the inter-cluster distance matrix.
4. Merge the nearest clusters.
5. Repeat steps 2 through 4 until all components are part of one cluster.

Minimum Spanning Tree Example

Program	CPU Time	Disk I/O
A	2	4
B	3	5
C	1	6
D	4	3
E	5	2

- Step 1: Consider five clusters with ith cluster consisting solely of ith program.
- Step 2: The centroids are {2, 4}, {3, 5}, {1, 6}, {4, 3}, and {5, 2}.

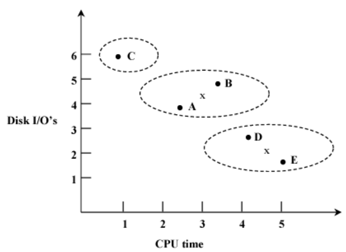
Spanning Tree Example (Cont)

- Step 3: The Euclidean distance is:

	Program				
Program	A	B	C	D	E
A	0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
B		0	$\sqrt{5}$	$\sqrt{5}$	$\sqrt{13}$
C			0	$\sqrt{18}$	$\sqrt{32}$
D				0	$\sqrt{2}$
E					0

- Step 4: Minimum inter-cluster distance = $\sqrt{2}$. Merge A+B, D+E.

Spanning Tree Example (Cont)



- Step 2: The centroid of cluster pair AB is {(2+3)÷ 2, (4+5) ÷ 2}, that is, {2.5, 4.5}. Similarly, the centroid of pair DE is {4.5, 2.5}.

Spanning Tree Example (Cont)

- Step 3: The distance matrix is:

	Program		
Program	AB	C	DE
AB	0	$\sqrt{4.5}$	$\sqrt{10.25}$
C		0	$\sqrt{24.4}$
DE			0

- Step 4: Merge AB and C.
- Step 2: The centroid of cluster ABC is {(2+3+1) ¥ 3, (4+5+6) ¥ 3}, that is, {2, 5}.

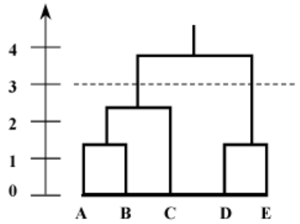
Spanning Tree Example (Cont)

- Step 3: The distance matrix is:

	Program	
Program	ABC	DE
ABC	0	$\sqrt{12.5}$
DE		0

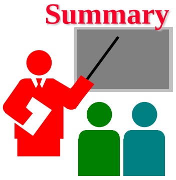
- Step 4: Minimum distance is 12.5.
  Merge ABC and DE $\Rightarrow$ Single Custer ABCDE

Dendogram



- Dendogram = Spanning Tree
- Purpose: Obtain clusters for any given maximum allowable intra-cluster distance.

Summary



- Workload Characterization = Models of workloads
- Averaging, Single parameter histogram, multi-parameter histograms, …
- Principal component analysis consists of finding parameter combinations that explain the most variation
- Clustering: divide workloads in groups that can be represented by a single benchmark