CS 747 Lecture 2: Mean Value Analysis (MVA)

Big Picture:

- A special class of queueing networks can be analyzed exactly using Mean Value Analysis
  - class is called "separable" or "product form" networks
  - examples include the Machine Repair Model and many other networks (including large, multi-class networks)
- A larger class of queueing networks can be analyzed approximately using heuristic (intuitive) extensions to MVA
- Key characteristics of MVA:
  - computes mean performance metrics
  - much more efficient than Markov Chains
  - based on the "Arrival Theorem" for separable networks

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

Definition of Single-Class Separable Networks:

1. Each queueing center must have one of the following scheduling disciplines:
   - First-Come First-Served (FCFS)
   - Processor Sharing (PS)
   - Last-Come First-Served Preemptive Resume (LCFSPR)
   - plus a few others that are not usually found in practice
2. If the scheduling discipline is **FCFS**,
   - the service times must be exponentially distributed
3. External arrivals (if any) must be Poisson

   (possibly state dependent)
4. Customer routing probabilities must be state-independent

Note:

- The above definition is equivalent to the less technical definition in LZGS section 6.5
- service times at PS, LCFSPR, and delay centers can have any distribution
- The above assumptions are stricter than the assumptions for ABA, except multi-server queues are allowed

Author: M. K. Vernon, *vernon@cs.wisc.edu*, 01/23/01

MVA - simple set of single-class model inputs:

K: number of queueing centers in the network

$\{D_k: 1 \le k \le K\}$: average demand at each queueing center

Z : sum of mean time at each delay center (e.g., mean "think" time)

λ or **N :** workload intensity

(i.e., arrival rate or fixed number of customers in the network)

- average demand at a queueing center is the average number of visits to the queue times the average service time per visit
- if the workload intensity varies, solve for each observed value (see Table 6.1 in LZGS)

MVA - model outputs:

X, R, Q,

X_k, R_k, Q_k, U_k

- $\mathbf{Q_k}$ is the average number of customers at center k
- $\mathbf{R_k}$ is the *total* average residence time at center **k**

  = $V_k \times$ average residence time per visit to **k**
- **R** may or may not include Z
- Distribution of queue length, $P[Q_k = n]$, $n=0,1,2,\ldots$, is not computed. Later in CS747 we will discuss a technique called convolution for computing queue length distributions.

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

MVA: Open, Single-Class Networks

$A_k(\lambda)$ = average number of customers in queue k at an arrival instant

For $\lambda < 1/D_{max}$ (i.e., $X = \lambda$):

$X_k(\lambda) = \lambda V_k$ [forced flow law]

$U_k(\lambda) = \lambda D_k$ [utilization law]

$R_k(\lambda) = V_k S_k [1 + A_k(\lambda)]$ (single-server queues)

$Q_k(\lambda) = \lambda R_k(\lambda)$ [Little's result]

$R(\lambda) = \Sigma R_k(\lambda)$, $Q(\lambda) = \Sigma Q_k(\lambda)$

- FCFS scheduling:

  avg. remaining service time at arrival instant = S_k
- LCFSPR scheduling: $A_k(\lambda)$ = avg. # that preempt
- PS scheduling: $A_k(\lambda)$ = avg. # that share the server
- We'll consider multi-server queues and other queue-length dependent service rates later [LZGS Chapters 8 & 20]

Arrival Theorem for open product-form networks (PFQNs):

$$A_k(\lambda) = Q_k(\lambda)$$

Since $A_k(\lambda) = Q_k(\lambda) = \lambda R_k(\lambda)$, for each queue:

$$R_k(\lambda) = D_k / [1 - U_k(\lambda)]$$

(each queue is solved separately; highly efficient solution)

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

Further remarks about the Arrival Theorem

$A_k(\lambda)$ = average number of customers in queue k at an arrival instant

- "in queue k" includes the customer in service
- "in queue k" doesn't include the arriving customer

Arrival Theorem for open single-class PFQNs:

$$A_k(\lambda) = Q_k(\lambda)$$

- one queue, no feedback: arrivals are Poisson, holds by PASTA
- multiple queues, feed forward network: by Burke's result arrivals to each queue are Poisson, holds by PASTA
- queueing network (one or more queues) with feedback: arrivals are not Poisson, but Arrival Theorem holds

  (surprising result, first proved by Jackson for open single-class networks in 1957; we will read Jackson's paper and generalizations later in cs747)

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

Applications of MVA, Open Networks

1) Consider a network that contains one single-server queue:

$$R(\lambda) = D / [1 - U(\lambda)]$$

- Result holds for: M/M/1, M/G/1 PS, or M/G/1 LCFSPR queue
- graph R vs U
  - at low utilization, $R \approx D$
  - impact of interference: $[1 - U(\lambda)]$ in denominator
  - curve qualitatively matches experience and intuition
- insights, due to convexity of R_k vs U_k:
  1) For a multi-queue network with a given D,
     a balanced system is optimal [balanced system: $D_k = D/K$]
  2) For a multi-queue network with a given throughput,
     an open system is pessimal
- $Q(\lambda) = U / (1-U)$

2) Networks with multiple queues:

$$U_k(\lambda) = \lambda D_k$$

$$R_k(\lambda) = D_k / [1 - U_k(\lambda)]$$

$$R = \Sigma R_k(\lambda)$$

- quick calculations (each queue is solved separately)
- solution time is O(K) (4-5 operations per queue)
- system modification analysis

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

MVA: Closed Separable Single-Class Networks

- MVA equations:

  Single-server queues: $R_k(N) = D_k [1 + A_k(N)]$

  $R(N) = \Sigma R_k(N)$

  $X(N) = N / [Z + R(N)]$

  $Q_k(N) = X(N) R_k(N)$

  $U_k(N) = X(N) D_k$

- Arrival Theorem: $A_k(N) = Q_k(N-1)$
  - $A_k(1) = 0 = Q_k(0)$
- solution algorithm:
  - $Q_k(0)$ are known
  - solve MVA equations for N = 1, then for N = 2, then 3,...
  - solution complexity: O(NK)
- Example: Table 6.2 in LZGS [go through this example]
- R vs N is convex:

  balanced system is optimal

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

Applications of Closed Separable Network Solution

1) Machine Repair Model, e.g., for office with 2 TAs, 1 workstation

2) Scherr's Ph.D. thesis [1965]:

question: can MRM accurately predict CTSS performance?

CTSS was an experimental interactive system

system swaps among user jobs, but only one active job at a time

Application of Machine Repair Model:

delay center = user think time

single server queue = avg processing time plus *disk service*

Results, Figure 6.2 in LZGS, R vs N:

predicted R and measured R are in good agreement

- very simple model, validated against actual system measurement
- separable model accurately predicted R even though the system violated an assumption for separable networks (i.e., service times were not exponentially distributed)

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01

Applications of Closed Separable Network MVA

3) Modification Analysis: Example in **LZGS** Section 6.3.2

- Batch system: one CPU, two I/O channels,
- 8 disks on one channel, 16 disks on the other
- Create QN; ignore overlapped i/o on third channel
- 3 proposed changes evaluated
- projected U_{cpu} for o.s. improvement plus disk upgrade was found to be accurate
- similar to ABA modification analysis, but more detailed estimates provide more precise evaluation of the particular proposed upgrades

4) Capacity planning: example in LZGS section 6.3.3

transaction workload; cpu, 40 disks

Author: M. K. Vernon, vernon@cs.wisc.edu, 01/23/01