Modeling Virtualized Environments in Simalytic® Models by Computing Missing Service Demand Parameters

CMG2009 Paper 9103, December 11, 2009

Dr. Tim R. Norton
Simalytic Solutions, LLC

719-635-5825
email: tim.norton@simalytic.com
http://www.simalytic.com

© 2009 Simalytic Solutions, LLC

OR

Using Virtual Data to Model Virtual Systems

© 2009 Simalytic Solutions, LLC
Introduction

Capacity Planning
- Capacity is Measured by Business Performance Objectives
  - Making decisions about resource requirements
- What do we have to buy and when do we have to buy it to make sure that the business applications perform at the level required to insure the business succeeds?
Introduction

Capacity Planning

- Two key aspects:
  - Demand for available resources
    - What do we have to buy
  - Effective completion of business work
    - When to buy it?
- Requires some predictive technique
  - Usually some form of model
    - Simple: trend or other statistical techniques
    - Advanced: simulation or queuing network

Virtualization

- Increasing parallelization within the host system
  - Increase productive business-related work
  - Increase the usage of resources
- Virtualized environment control program
  - Hypervisor
    - Usually implies a hardware implementation
  - VMM (Virtual Machine Monitor)
    - VMM often implies a software implementation
    - Will use VMM for all virtualization control programs
虚拟能型モデル

Server 1
VM1  VM2  VM3

Server 2
VM4  VM5  VM6

ネットワーク

仮想化

從來から

システムA  システムB  システムC

良いシステムレベルおよびプロセスレベルの測定

到達

システムA  システムB  システムC

システムA  システムB  システムC

システムレベルの測定
The Problem

◆ Measurements in Virtualization Environments
  ● Guest Operating Systems
    ▪ Most not Virtualization aware
      ■ Incorrect accounting for time when guest VM not active
        ◆ Rate based measurements incorrect
        ◆ Count based measurements valid
  ● Hypervisor
    ▪ Accurately measures Guest VM active time
    ▪ Cannot measure processes within the Guest O/S
  ● Affects measurements needed for models
    ■ VM System Level Service Times good
    ■ Process Level Service Times incorrect

Capacity Models

◆ What Virtualization does to Models
  ● Reduces accuracy – Measurement issues:
    ▪ System clock
    ▪ Accounting for dispatch time of other VMs
    ▪ Interrupts – delays and re-driven to VM
    ▪ VM vs. process priority
    ▪ Delays – where are they accounted for
    ▪ Virtualization overhead
      ■ VMM (Virtual Machine Monitor) and VM context switch
      ■ Interference from other VMs
  ● Complicates Workload Characterization
    ■ Shared resources
Where’s the Data

◆ If It’s Not There…
  ○ It’s missing!
◆ If It’s Wrong…
  ○ It’s missing!
◆ If You Can’t Trust It…
  ○ It’s missing!

A Proposed Solution

◆ Use known good measurements
  ○ VM utilization
    ■ from the VMM (Virtual Machine Monitor)
  ○ Transaction arrivals
    ■ from application measurements
  ○ Transaction response times
    ■ from application measurements
◆ Use a Simalytic Model
  ■ Builds relationships
  ■ Iterative design
  ■ Leverages existing tools
Virtualization Metrics

VM – Guest O/S Measurements

Needed to model applications

VMM (Known Good) Measurements

Model Requirements

- Application Transaction Arrivals
  - Count of transactions over many intervals

- Application Response Times
  - Always needed for validation of model results
    - Now needed for Solver calculations

- Service Demand
  - Menascé technique to compute class service demand from total service demand
    - VMM measurements used to compute VM level measurements that guest O/S cannot provide
Implementation Approach

- **Collect Measurement Data**
  - Application – counts and response times
  - VMM resource usage

- **Compute Service Demand by Application**
  - Validate against measured response times

- **Build Model of Overall Virtualization Environment**
  - Using a Simalytic Model to express the relationships between applications

Implementation

- **Measurements**
  - Collect measurements
    - From the applications
      - Actual response times
      - Actual transaction counts
    - From the VMM
      - Total utilization for each resource (CPU, disks, etc.)
  - Multiple intervals
    - Measurements are needed for many intervals
      - Variety is important!
        - but avoid problem areas like low utilization effect
          - Low to high counts for each application
          - Low to high resource utilization
Implementation

◆ **Compute Service Demand**
  - Use collected measurements to construct multiple Open Multiclass QN formulae
    - Each uses the transaction response time and count for each class (application) along with the total service demand for each resource (CPU, disks, etc.) for one measurement interval.
  - Solve the non-linear constraint problem
    - Can be done with Microsoft Excel Solver

◆ **Model Beyond What was Measured**
  - Measurement data provides historical view
    - Many intervals available
      - But at lower than expected future traffic volume
  - Model results provide future view
    - Answer the classic questions:
      - When does response time become unacceptable?
      - What resource saturates first?
  - The following example takes this approach
    - Measure system and applications at lower volumes
    - Predict behavior at higher volumes
      - Different than example in the paper
Excel Solver Spreadsheet

- Solves Multiple Equations
  - Initial Guess and Solver Results
  - Known Good Measurements
  - “Actual” measurements
  - Computed measurements
  - Validation

Using Excel Solver

- Solver Results – Example Spreadsheets
  - Start with initial guess
    - Same for both classes – calculated from VMM resource measurements divided sum of trans in both classes.
  - Computes values for both classes
    - non-linear constraints
  - Actual Service Demand Data
    - Not real measurements
      - Validate approach
      - Generate data for testing
Using Excel Solver

◆ “Actual” Service Demand Values
  ○ Used at low transaction volume to create known good response times and device utilizations
    ◦ Used for Solver goals
    ◦ Simulates actual known good measurements
      • M/M/1 formulae

Using Excel Solver

◆ “Actual” Service Demand Values
  ○ Used at high transaction volume to create “actual” (projected) response times and device utilizations
    ◦ Predicts results at higher arrival rates
    ◦ Simulates future actual measurements to validate Solver results
Using Excel Solver

- **Computed Service Demand Values**
  - Used to create “computed” response times and device utilizations
    - Solver results
    - Validated against “projected” values

- **Solver Answer Report**
  - Shows details behind solution
Application Performance

- Two groups of measurements
  - "Actuals" are data used to create the models
  - "Projections" are data used as post-model measurements
  - Each class (application) shows response time increase as utilizations go up

Application Prediction

- How Well do Computed Results Match
  - Actual Measurements
  - Results using Computed Service Demands
  - Some difference but same trend
Usage

◆ How Can Computed Service Demands be Used?

○ Stand-alone models
  ■ Same as measured service demands
    ● Adjustments may be needed for VMM overhead and other interference – similar to other model calibrations

○ Simalytic Models
  ■ Enhanced Simalytic Function for multi-tier models
    ● Dynamic calculations to
      ◆ simulate complex usage patterns
      ◆ account for effects of spikes in other workloads

Conclusion

◆ Future Work

○ Explore criteria around measurement collection
  ■ How many intervals needed for minimal effectiveness
  ■ What improves accuracy

○ Incorporate solver into modeling tool
  ■ Possibly as enhanced Simalytic Function
Conclusion

◆ Valid Approach
  ○ Works with synthetic test data
    ■ Some differences
      ● between projected and computed results
    ■ But trends usable for planning
  ○ Refinement needed
    ■ account for VMM overhead
  ○ Model virtual systems with virtual data

Missing Data Doesn’t Stop a Real Modeler!

Presentation and spreadsheet will be available at: http://www.simalytic.com/Papers.html