The Effect of Workload Groupings on Distributed Transaction Capacity Models

Dr. Tim R. Norton
Simalytic Solutions, LLC
Colorado Springs, CO
www.simalytic.com
tim.norton@simalytic.com

This Study

Why Model?

- "The Objective"
- "The Problem"
- "The Solution"
- "The Choices"

Understanding the Choices

- The Model
- The Conclusion

"The Objective"

- Analysis:
  - Resource Usage
    - What is our current bottleneck?
  - User Response
    - Are we losing business because Xxxx takes too long?
- Prediction:
  - Growth
    - When do we need a bigger …?
  - Change
    - What happens if we …?

"The Problem"

- Too Much Detail
  - Logging individual transactions
  - Many resources (processes, disks, networks, etc.)
  - Many time intervals (15 min = 672 / week; 5 min = 2016 / week)
- Technology vs. Business
  - Hard to map business usage to application design
  - Relating interval and event based measurements
- Reuse
  - Different business functions use the same utility transactions

"The Solution"

- Aggregation:
  - Group transactions together
    - by size, response time, name, location, type, …
  - Group resources together
    - by type, response time, location, name, ...
  - Group time intervals together
    - by length (week, month), type (shift, season), …
  - Group business functions together
    - by name, function, application, business unit, …
- Proration:
  - Split utility workloads across functional workloads
    - by percentage, counts, guesses, …

"The Choices"

- What is the Objective?
  - Analysis?
  - Prediction?
  - Technical?
  - Business?
- How to Pick the Groups?
  - Resource usage or Business usage view?
  - Short term or long term?
  - Quick and dirty or detailed and precise?

Study Model Goals

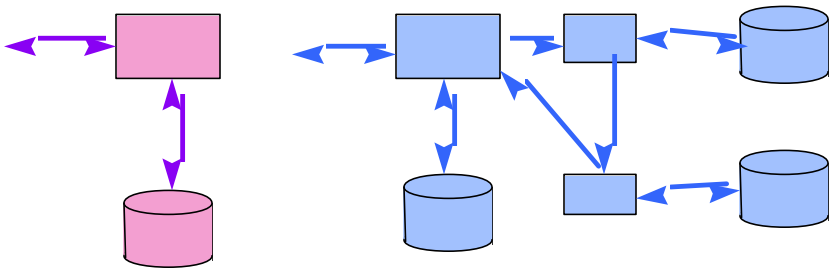
- Investigate the effects in a model of using workloads with different groupings of the same transactions.
- Investigate the effects of moving workloads from a single system with slow resources to a distributed environment with faster resources and network delays.

Study Overview

- Hypothetical Transaction Data
- Simulated Groupings for Workloads
  - Business Function Groupings
  - Resource Usage Groupings
- Two Scenarios
  - Local processing only
  - Local and network processing
- Open Queuing Network Model

The Scenarios

- #1 Slow CPU and Disks but no Network
- #2 Fast CPU and Disks over the Network



The Scenario Service Times

- #1 Slow CPU and Disks but no Network
  - CPU 0.0200
  - Disk 1 0.0120
  - Disk 2 0.0230
  - Disk 3 0.0350
  - Disk 4 0.0410
  - Network 0.0000
- #2 Fast CPU and Disks over the Network
  - CPU 0.0010
  - Disk 1 0.0120
  - Disk 2 0.0100
  - Disk 3 0.0030
  - Disk 4 0.0040
  - Network 0.2800

The Transactions

Partial List of Transaction Data

Transaction Name	Workload Groups 1	Workload Groups 2	Workload Groups 3	Workload Groups 4	Workload Groups 5	Service Time	CPU Units	Disk 1 Units	Disk 2 Units	Disk 3 Units	Disk 4 Units	Network Units	Transaction Count
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a	A	A1	Aa	a	1	1.7	9	84	64	9	12	0	6
A1a Average						1.7	9	84	64	9	12	0	6
A1b	A	A1	Ab	b	1	9.9	9	84	64	225	12	27	4
A1b	A	A1	Ab	b	1	9.9	9	84	64	225	12	27	4
A1b	A	A1	Ab	b	1	9.9	9	84	64	225	12	27	4
A1b	A	A1	Ab	b	1	9.9	9	84	64	225	12	27	4
A1b Average						9.9	9	84	64	225	12	27	4
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c	A	A1	Ac	c	1	22.9	9	84	64	225	720	63	7
A1c Average						22.9	9	84	64	225	720	63	7
A2a	A	A2	Aa	a	2	0.7	36	21	36	9	12	0	5
A2a	A	A2	Aa	a	2	0.7	36	21	36	9	12	0	5
A2a	A	A2	Aa	a	2	0.7	36	21	36	9	12	0	5
A2a	A	A2	Aa	a	2	0.7	36	21	36	9	12	0	5
A2a	A	A2	Aa	a	2	0.7	36	21	36	9	12	0	5
A2a Average						0.7	36	21	36	9	12	0	5

Transaction Profiles

	Volume	CPU	Disk 1	Disk 2	Disk 3	Disk 4	Network
Axx	H						
Bxx	M						
Cxx	L						
x1x		L	H	H			
x2x		M	M	M			
x3x		H	L	L			
xxa					L	L	none
xxb					H	L	M
xxc					H	H	H
	H=high, M=medium, L=low usage of the resourse or relative volume						

Work Groups

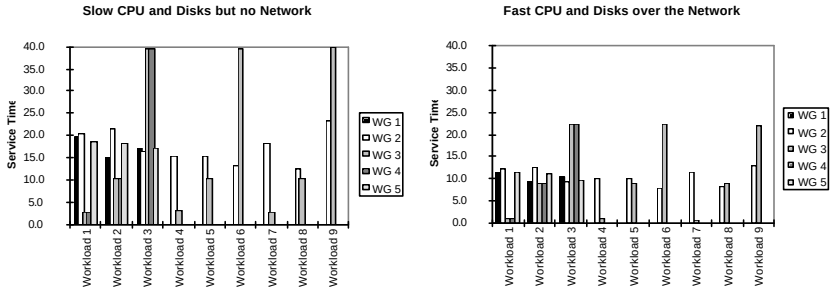
- Five Work Groups
  - Provides a name for each group of workloads
  - Each with a different transaction mix
  - Three or nine Workloads per WG
- Workloads in Each Work Group (WG)
  - WG1: A?? B?? C??
  - WG2: A1? A2? A3? B1? B2? B3? C1? C2? C3?
  - WG3: A?a A?b A?c B?a B?b B?c C?a C?b C?c
  - WG4: ??a ??b ??c
  - WG5: ?1? ?2? ?3?

Work Groups

- Perspective
  - Each WG provides a different view into the same transaction data to meet different modeling objectives.
- Examples:
  - WG1: A?? B?? C??
    - Three Workloads based solely on transaction volumes
  - WG2: A1? A2? A3? B1? B2? B3? C1? C2? C3?
    - Nine Workloads based transaction volume, CPU and some disks
  - WG4: ??a ??b ??c
    - Three Workloads based some disks and network usage

Service Times

- Scenario #1
- Scenario #2



The Model Used

- Queuing Network
  - Multiple classes (workloads) in each Work Group
  - Open queuing network (unlimited source of trans)
  - Load independent servers – service time scenarios:
    1: slow CPU, all local slow disks and no network
    2: fast CPU, local and disk subsystems, network
- Program OPENQN.EXE

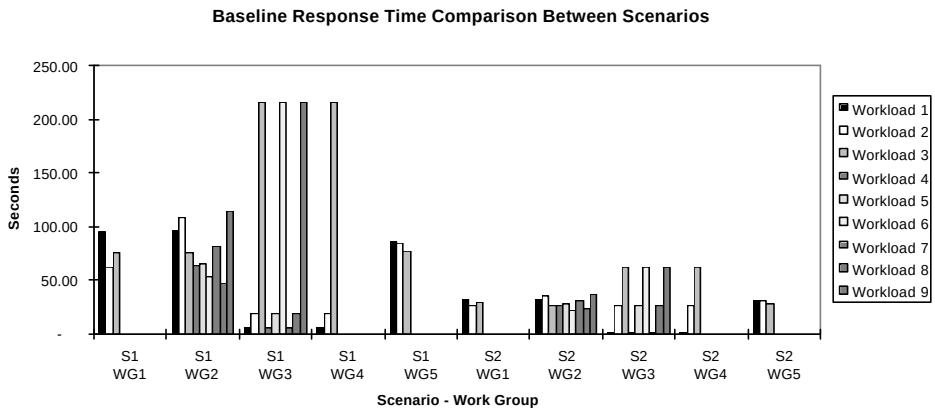
Capacity Planning and Performance Modeling: from mainframes to client-server systems,
by D. Menascé, V. Almeida, and L. Dowdy
Published by Prentice Hall, 1994.

What Was Modeled

- Baseline
  - Each Work Group for each Scenario
- Growth
  - Four new periods
  - Each transaction had different growth
  - Workload growth depends on transaction mix
    - Results would be different if the growth rates were constant for each Workload in the different Work Groups (as would be the case for strickly business unit aggregation).

 Rocky Mountain CMG, April 3, 2003 17

Baseline Response Times



 Rocky Mountain CMG, April 3, 2003 18

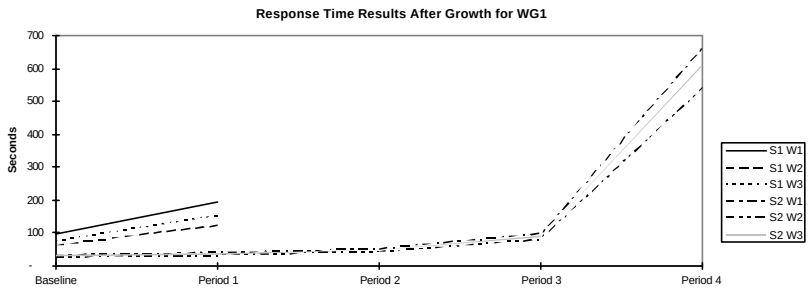
Overall Workload Growth

- Ratio of Period 4 to Baseline (P4/B)

	WG1 X??	WG2 X9?	WG3 X?x	WG4 ??x	WG5 ?9?
Workload 1	1.65	1.66	1.64	1.41	1.43
Workload 2	1.12	1.64	1.63	1.34	1.38
Workload 3	1.02	1.67	1.68	1.48	1.43
Workload 4		1.11	1.11		
Workload 5		1.12	1.14		
Workload 6		1.12	1.08		
Workload 7		1.02	1.02		
Workload 8		1.01	1.01		
Workload 9		1.02	1.03		

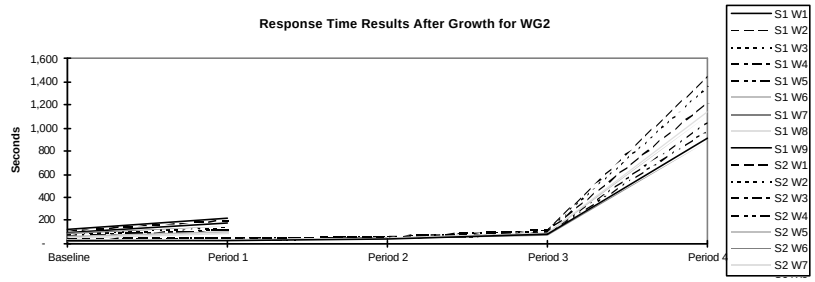
 Rocky Mountain CMG, April 3, 2003 19

Work Group 1 Growth



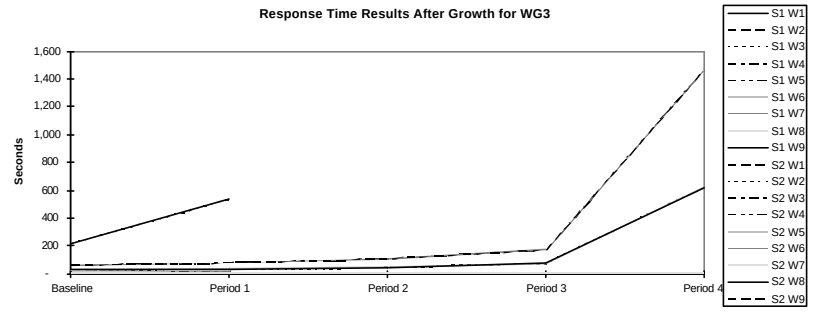
 Rocky Mountain CMG, April 3, 2003 20

Work Group 2 Growth



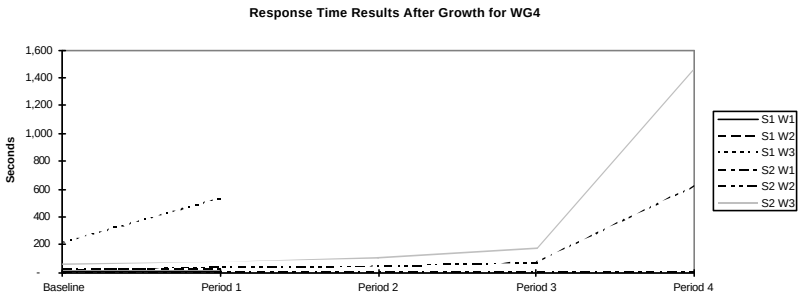
 Rocky Mountain CMG, April 3, 2003 21

Work Group 3 Growth



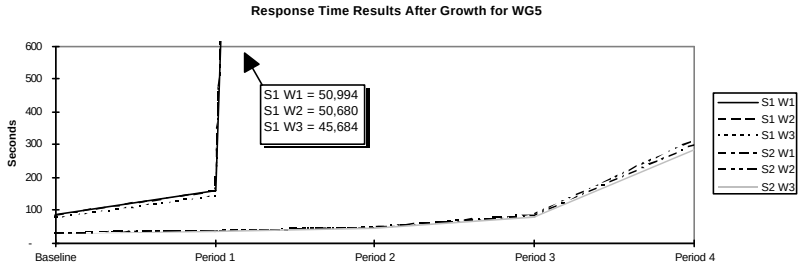
 Rocky Mountain CMG, April 3, 2003 22

Work Group 4 Growth



 Rocky Mountain CMG, April 3, 2003 23

Work Group 5 Growth



 Rocky Mountain CMG, April 3, 2003 24

Conclusions

- Expected Queuing Effects
  - More pronounced with slower servers
    - Service times still limit volumes
  - "Knee of the Curve"
    - Sudden change in response time with a small change in arrival rate
- Local vs. Remote
  - Network delay can be offset by lower service times and more servers
    - Reduces queuing by spreading out the number of transactions in the system across more queues.

Conclusions

- Work Groups
  - Grouping has major impact
    - Some groups cluster results (WG3 & WG4)
    - Some groups spread results (WG2 & WG5)
    - Some increase response time differences (WG3 & WG4)
    - Some decrease response time differences (WG5)
  - Must reflect growth as well as initial mix
    - Workloads that are homogeneous in one relationship (volume, resource usage, etc.) are not necessarily homogeneous in another (growth).
    - Which is more important: a better "average" transaction to start with or a growth rate closer to reality?

Conclusions

- Work Groups
  - Workload resource bottleneck sensitivity
    - An overlooked metric of homogeneousness
    - Often more important that other metrics
    - Difficult to quantify without a lot of work
    - Identifiable when results driven to extremes with none in-between
- Back to the Objective
  - Which is better?
    - A model that validates well (accurate baseline)
    - A model that predicts well (accurate future performance)

Questions

?