End-To-End Scaling: The Response Time Pipe

CMG2001 Session 3208, December 4, 2001

http://www.simalytic.com/CMG01/3208ppt.pdf

Dr. Tim R. Norton

Simalytic Solutions, LLC

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

Agenda

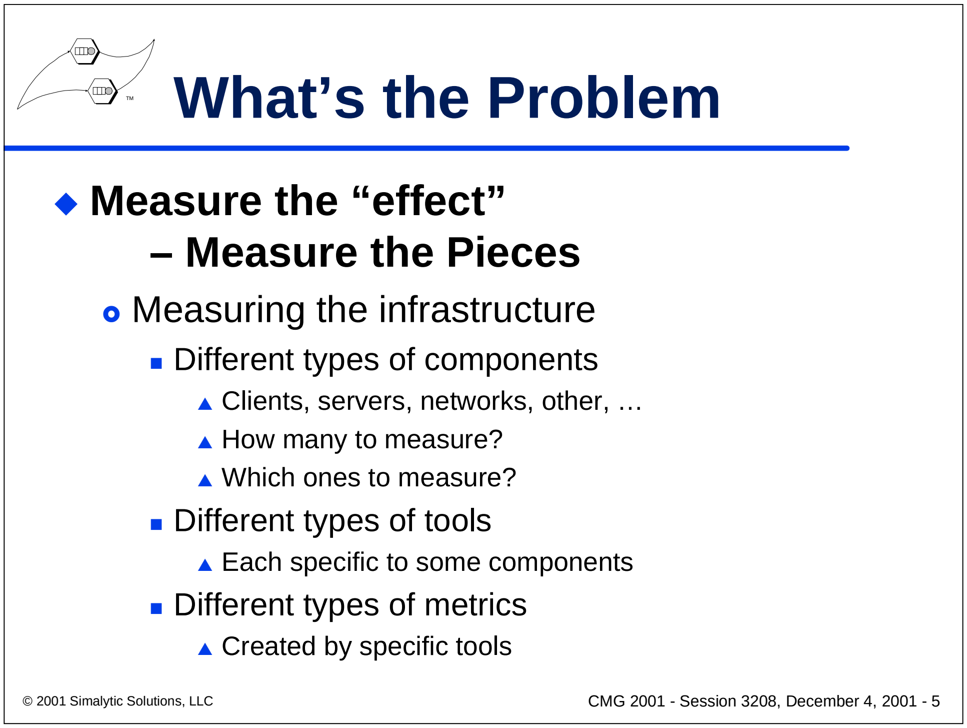
- **What's the Problem**
  - Background
- **The Response Time Pipe Solution**
  - Techniques that fit the problem
- **Response Time Pipe Example**
  - Sample solution to a hypothetical situation

What’s the Problem

- **How does the performance of a computer application effect the business?**
  - Defining the relationship between the two:
    - The business result when the application changes
    - The application result when the business changes
  - What is the “effect”?
    - Requires measuring both
  - Implies there is a “good” and a “bad”
    - Assessment of the relationship
    - How to predict when it will become “bad”?
  - How to use performance numbers to answer business (i.e., financial) questions?

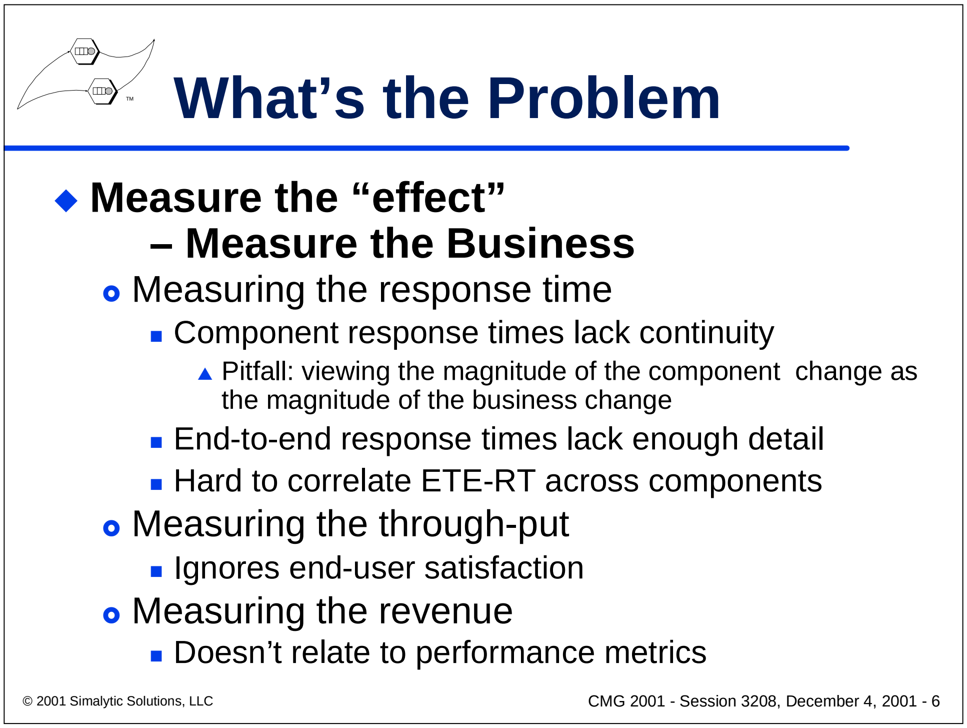
What’s the Problem

- **Measure the “effect”**
  – Measure the Pieces
  - Measuring the application
    - Different types of applications
      - Fat/thin client, multi-tier, web based, proprietary, …
    - Different units of work
      - Transactions, messages, interactive, asynchronous, …
    - What is the end-user’s experience?
    - Measure everything or just what’s “important”?



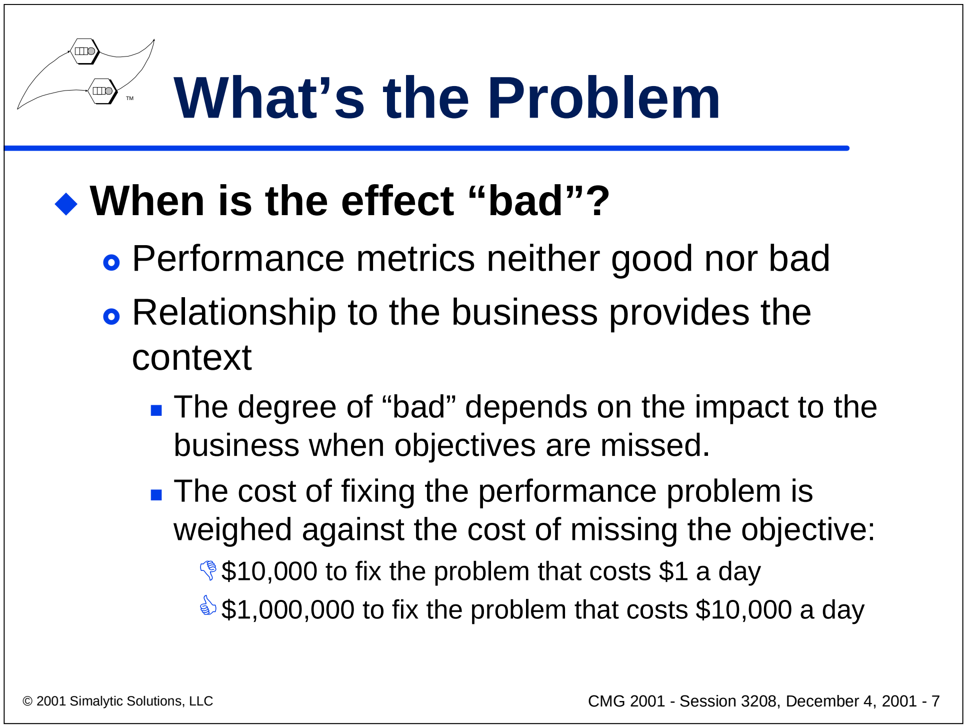
What’s the Problem

- **Measure the “effect”**
  - **– Measure the Pieces**
  - Measuring the infrastructure
    - Different types of components
      - Clients, servers, networks, other, …
      - How many to measure?
      - Which ones to measure?
    - Different types of tools
      - Each specific to some components
    - Different types of metrics
      - Created by specific tools



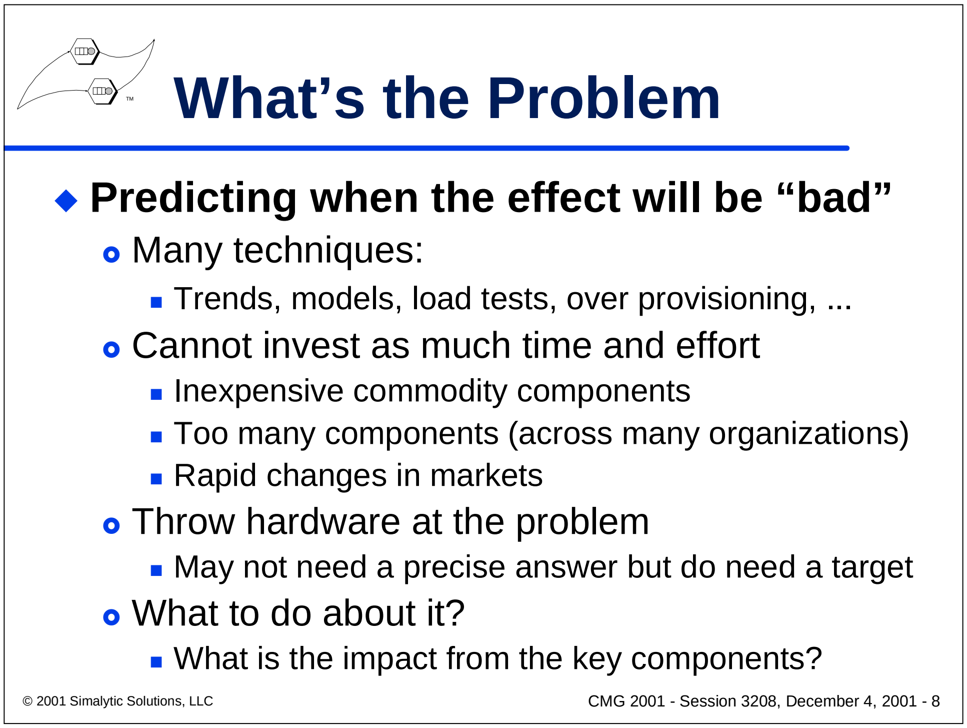
What’s the Problem

- **Measure the “effect”**
  - **– Measure the Business**
  - Measuring the response time
    - Component response times lack continuity
      - Pitfall: viewing the magnitude of the component change as the magnitude of the business change
    - End-to-end response times lack enough detail
    - Hard to correlate ETE-RT across components
  - Measuring the through-put
    - Ignores end-user satisfaction
  - Measuring the revenue
    - Doesn’t relate to performance metrics



What’s the Problem

- When is the effect “bad”?
  - Performance metrics neither good nor bad
  - Relationship to the business provides the context
    - The degree of “bad” depends on the impact to the business when objectives are missed.
    - The cost of fixing the performance problem is weighed against the cost of missing the objective:
      - 👎 $10,000 to fix the problem that costs $1 a day
      - 👍 $1,000,000 to fix the problem that costs $10,000 a day



What’s the Problem

- Predicting when the effect will be “bad”
  - Many techniques:
    - Trends, models, load tests, over provisioning, ...
  - Cannot invest as much time and effort
    - Inexpensive commodity components
    - Too many components (across many organizations)
    - Rapid changes in markets
  - Throw hardware at the problem
    - May not need a precise answer but do need a target
  - What to do about it?
    - What is the impact from the key components?

What's the Problem

What's Needed in a Solution?

- Need an approximation technique
  - Easy to use without years of experience
  - Identifies areas of concern
  - Eliminates areas that don't matter (right now)
  - Usable results quickly enough for business decisions
- Need a technique to tie all the measurement pieces together, regardless of sources
- Need a technique to relate the overall result to the business but still identify key components
  - Provides focus for existing analysis techniques

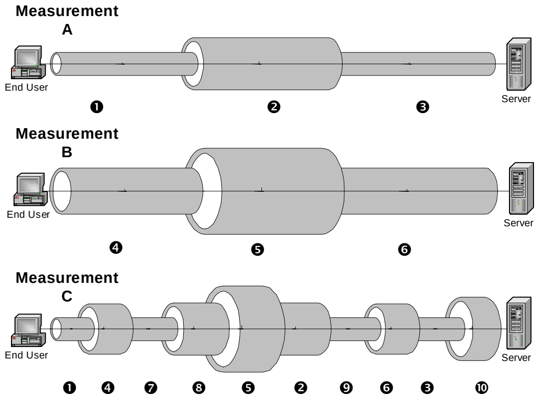
Response Time Pipe Solution

What is a Response Time Pipe?

- Way to visualize the relationships between components used by an application.
- A technique that quickly connects different types of component performance measurements or approximations.
- A technique to relate the performance of the components to the business objective.

Response Time Pipe Solution

- **Why a Pipe?**
  - To provide a visual framework that expresses:
    - Capacity
    - Connection
    - Flow
    - Sections
    - Constrictions



  - Looking at different sections provides different perceptions of capacity and performance

Response Time Pipe Solution

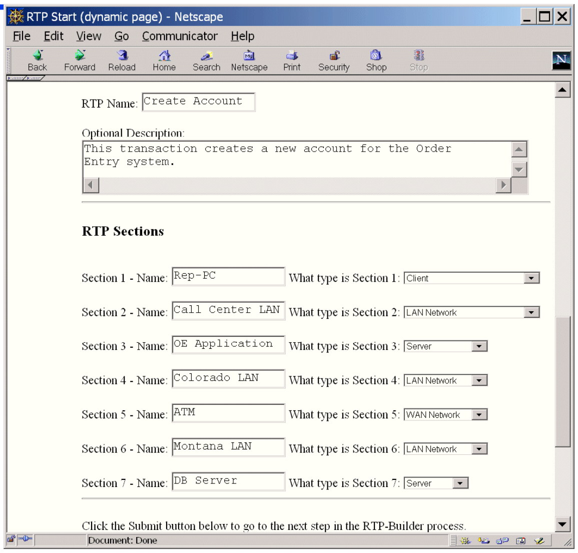
- **How to Build an RTP**
  - Identify a unit of business work (transaction)
  - Establish the overall objective
  - Measure the overall response time
  - Divide the infrastructure into sections
  - Identify the transaction flow across the sections
  - Measure each section with appropriate metrics
  - Map the metrics to transaction response times
  - Connect the response times from all sections

Response Time Pipe Example

- **Hypothetical Situation and Infrastructure**
  - Operators service customers in a call center
  - Simple *Create Account* Transaction
  - Multi-tier infrastructure
    - Client PC
    - Call Center LAN
    - Order Entry Application Server
    - Network segments (LAN→WAN→LAN)
    - Database Sever

Response Time Pipe Example

- **Define each section of the RTP**
  - Name
  - Type of section
    - Client
    - Server
    - LAN
    - WAN



RTP Name: Create Account

Optional Description:
This transaction creates a new account for the Order Entry system.

RTP Sections

Section 1 - Name: Rep-PC What type is Section 1: Client

Section 2 - Name: Call Center LAN What type is Section 2: LAN Network

Section 3 - Name: OE Application What type is Section 3: Server

Section 4 - Name: Colorado LAN What type is Section 4: LAN Network

Section 5 - Name: ATM What type is Section 5: WAN Network

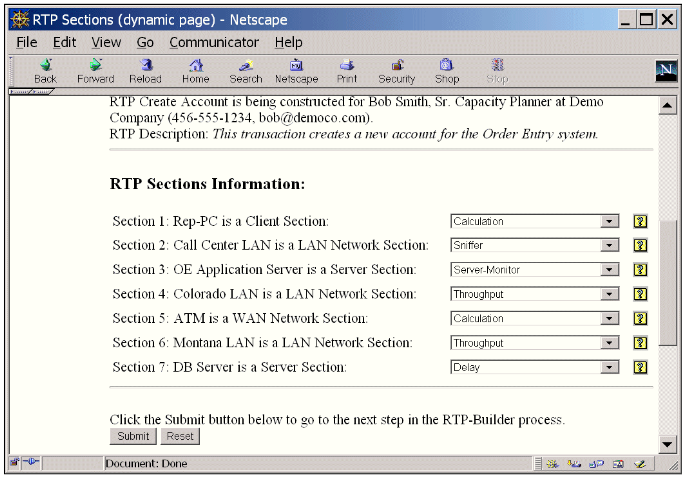
Section 6 - Name: Montana LAN What type is Section 6: LAN Network

Section 7 - Name: DB Server What type is Section 7: Server

Click the Submit button below to go to the next step in the RTP-Builder process.

Response Time Pipe Example

- **Define how each section is measured**
  - Calculated
  - Sniffer
  - Monitor
  - Through-put
  - Delay



RTP Sections (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

RTP Create Account is being constructed for Bob Smith, Sr. Capacity Planner at Demo Company (456-555-1234, bob@democo.com).
RTP Description: *This transaction creates a new account for the Order Entry system.*

RTP Sections Information:

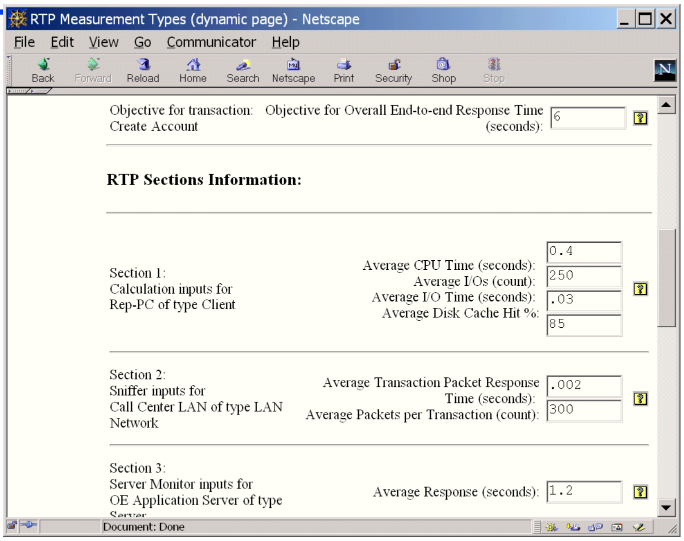
Section 1: Rep-PC is a Client Section:	Calculation
Section 2: Call Center LAN is a LAN Network Section:	Sniffer
Section 3: OE Application Server is a Server Section:	Server-Monitor
Section 4: Colorado LAN is a LAN Network Section:	Throughput
Section 5: ATM is a WAN Network Section:	Calculation
Section 6: Montana LAN is a LAN Network Section:	Throughput
Section 7: DB Server is a Server Section:	Delay

Click the Submit button below to go to the next step in the RTP-Builder process.

Submit Reset

Response Time Pipe Example

- **Overall objective**
- **Enter the transaction measures for each section**
  - Client calc: CPU & I/O
  - Sniffer: Packet count and response time
  - Monitor: measured value



RTP Measurement Types (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

Objective for transaction: Create Account — Objective for Overall End-to-end Response Time (seconds): 6

RTP Sections Information:

Section 1:
Calculation inputs for
Rep-PC of type Client

Average CPU Time (seconds): 0.4
Average I/Os (count): 250
Average I/O Time (seconds): .03
Average Disk Cache Hit %: 85

Section 2:
Sniffer inputs for
Call Center LAN of type LAN Network

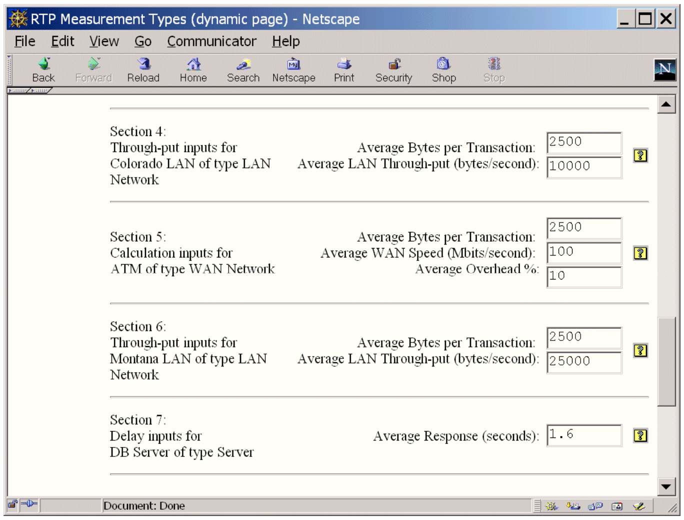
Average Transaction Packet Response Time (seconds): .002
Average Packets per Transaction (count): 300

Section 3:
Server Monitor inputs for
OE Application Server of type Server

Average Response (seconds): 1.2

Response Time Pipe Example

- **Enter the transaction measures for each section**
  - Through-put: bytes and through-put
  - WAN calc: bytes, speed and over-head
  - Delay: value



Section 4:
Through-put inputs for
Colorado LAN of type LAN Network

Average Bytes per Transaction: 2500
Average LAN Through-put (bytes/second): 10000

Section 5:
Calculation inputs for
ATM of type WAN Network

Average Bytes per Transaction: 2500
Average WAN Speed (Mbits/second): 100
Average Overhead %: 10

Section 6:
Through-put inputs for
Montana LAN of type LAN Network

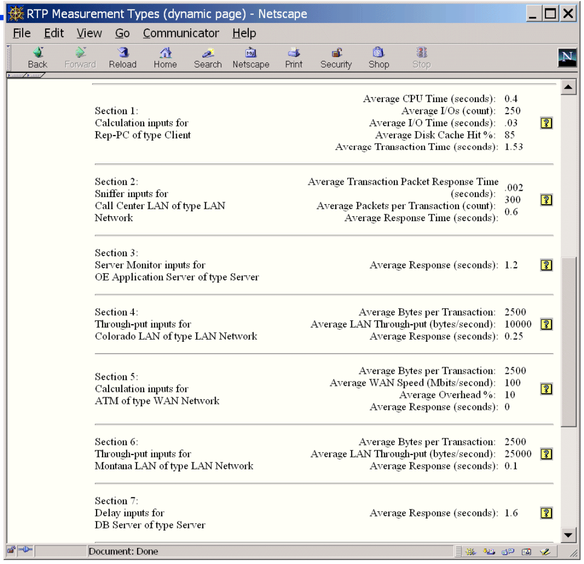
Average Bytes per Transaction: 2500
Average LAN Through-put (bytes/second): 25000

Section 7:
Delay inputs for
DB Server of type Server

Average Response (seconds): 1.6

Response Time Pipe Example

- **Calculate the transaction response times for each section**
  - Calc: add the component times
  - Sniffer: packet response time * count
  - Monitor: value
  - Through-put: based on total bytes
  - Delay: value



Section 1:
Calculation inputs for
Rep-PC of type Client

Average CPU Time (seconds): 0.4
Average I/Os (count): 250
Average I/O Time (seconds): .03
Average Disk Cache Hit %: 85
Average Transaction Time (seconds): 1.53

Section 2:
Sniffer inputs for
Call Center LAN of type LAN Network

Average Transaction Packet Response Time (seconds): .002
Average Packets per Transaction (count): 300
Average Response Time (seconds): 0.6

Section 3:
Server Monitor inputs for
OE Application Server of type Server

Average Response (seconds): 1.2

Section 4:
Through-put inputs for
Colorado LAN of type LAN Network

Average Bytes per Transaction: 2500
Average LAN Through-put (bytes/second): 10000
Average Response (seconds): 0.25

Section 5:
Calculation inputs for
ATM of type WAN Network

Average Bytes per Transaction: 2500
Average WAN Speed (Mbits/second): 100
Average Overhead %: 10
Average Response (seconds): 0

Section 6:
Through-put inputs for
Montana LAN of type LAN Network

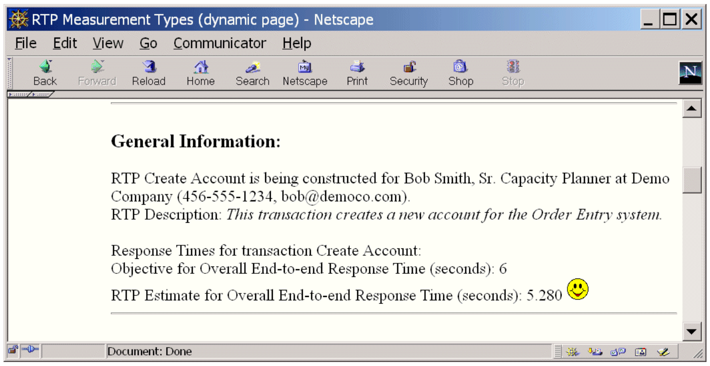
Average Bytes per Transaction: 2500
Average LAN Through-put (bytes/second): 25000
Average Response (seconds): 0.1

Section 7:
Delay inputs for
DB Server of type Server

Average Response (seconds): 1.6

Response Time Pipe Example

- **Compare the estimate to the objective**
  - First indicator of “goodness” or “badness”
    - “Best case” estimate of transaction response time


RTP Measurement Types (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

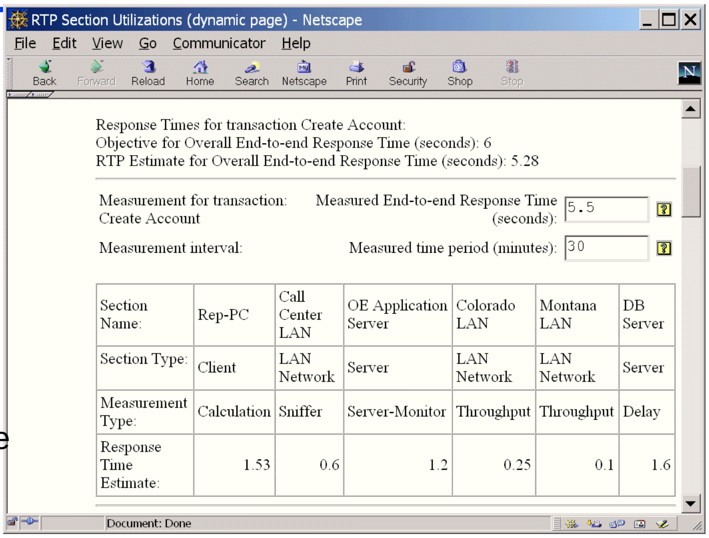
General Information:

RTP Create Account is being constructed for Bob Smith, Sr. Capacity Planner at Demo Company (456-555-1234, bob@democo.com).
RTP Description: *This transaction creates a new account for the Order Entry system.*

Response Times for transaction Create Account:
Objective for Overall End-to-end Response Time (seconds): 6
RTP Estimate for Overall End-to-end Response Time (seconds): 5.280 ☺

Response Time Pipe Example

- **Build the pipe**
  - Each section
  - Type
  - How it’s measured
  - Response times
- **Measured:**
  - Overall response time
  - Interval


RTP Section Utilizations (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

Response Times for transaction Create Account:
Objective for Overall End-to-end Response Time (seconds): 6
RTP Estimate for Overall End-to-end Response Time (seconds): 5.28

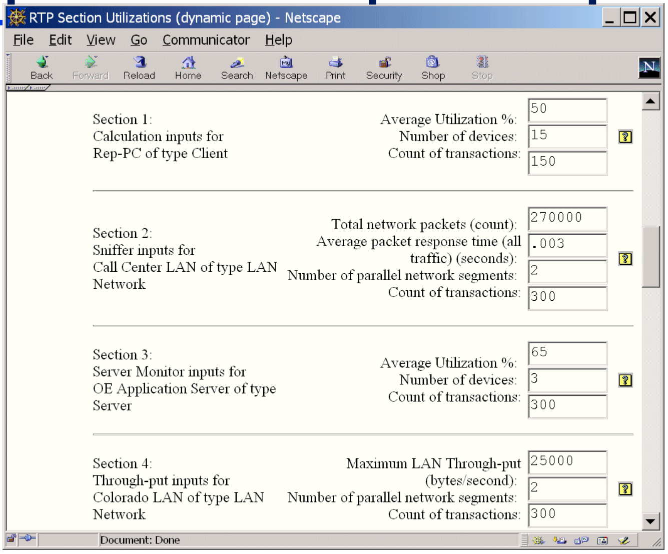
Measurement for transaction: Create Account — Measured End-to-end Response Time (seconds): 5.5 ?

Measurement interval: — Measured time period (minutes): 30 ?

Section Name:	Rep-PC	Call Center LAN	OE Application Server	Colorado LAN	Montana LAN	DB Server
Section Type:	Client	LAN Network	Server	LAN Network	LAN Network	Server
Measurement Type:	Calculation	Sniffer	Server-Monitor	Throughput	Throughput	Delay
Response Time Estimate:	1.53	0.6	1.2	0.25	0.1	1.6

Response Time Pipe Example

- Add current load information
  - utilizations
  - transaction counts
  - packet counts
  - byte counts
  - parallelism


RTP Section Utilizations (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop

Section 1:
Calculation inputs for
Rep-PC of type Client

Average Utilization %: 50
Number of devices: 15
Count of transactions: 150

Section 2:
Sniffer inputs for
Call Center LAN of type LAN Network

Total network packets (count): 270000
Average packet response time (all traffic) (seconds): .003
Number of parallel network segments: 2
Count of transactions: 300

Section 3:
Server Monitor inputs for
OE Application Server of type Server

Average Utilization %: 65
Number of devices: 3
Count of transactions: 300

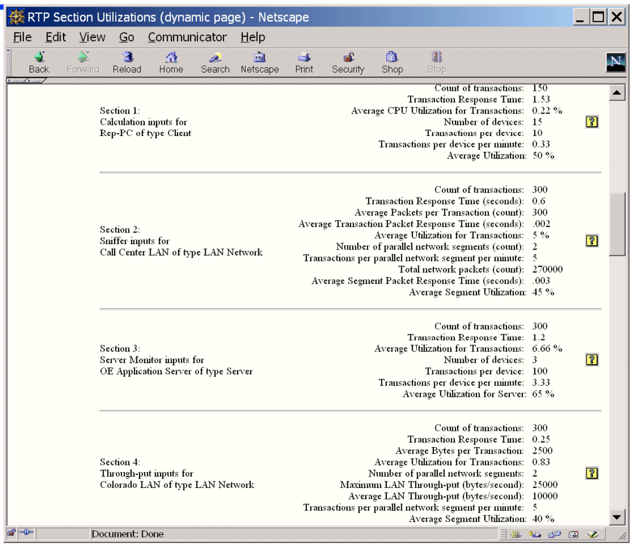
Section 4:
Through-put inputs for
Colorado LAN of type LAN Network

Maximum LAN Through-put (bytes/second): 25000
Number of parallel network segments: 2
Count of transactions: 300

Document: Done

Response Time Pipe Example

- Calculations for each section
  - New transaction response times
  - Transaction workload utilization
  - Overall utilization
  - Accounts for effect of current load


RTP Section Utilizations (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop

Section 1:
Calculation inputs for
Rep-PC of type Client

Count of transactions: 150
Transaction Response Time: 1.53
Average CPU Utilization for Transactions: 0.22 %
Number of devices: 15
Transactions per device: 10
Transactions per device per minute: 0.33
Average Utilization: 50 %

Section 2:
Sniffer inputs for
Call Center LAN of type LAN Network

Count of transactions: 300
Transaction Response Time (seconds): 0.6
Average Packets per Transaction (count): 300
Average Transaction Packet Response Time (seconds): .002
Average Utilization for Transactions: 5 %
Number of parallel network segments (count): 2
Transactions per parallel network segment per minute: 5
Total network packets (count): 270000
Average Segment Packet Response Time (seconds): .003
Average Segment Utilization: 45 %

Section 3:
Server Monitor inputs for
OE Application Server of type Server

Count of transactions: 300
Transaction Response Time: 1.2
Average Utilization for Transactions: 6.66 %
Number of devices: 3
Transactions per device: 100
Transactions per device per minute: 3.33
Average Utilization for Server: 65 %

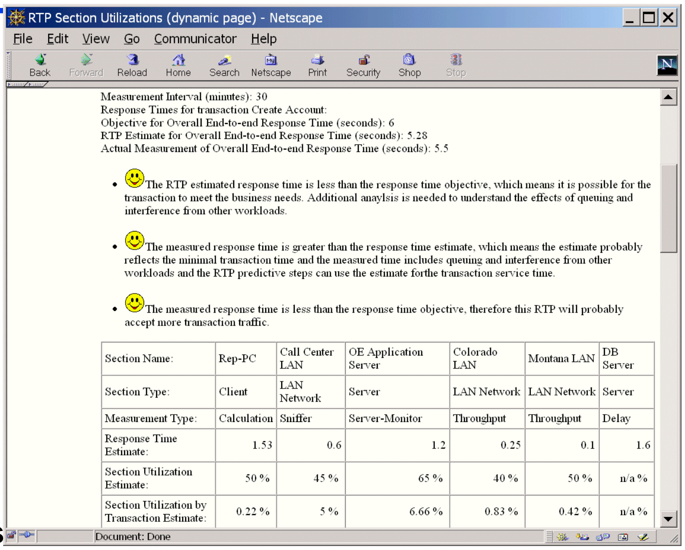
Section 4:
Through-put inputs for
Colorado LAN of type LAN Network

Count of transactions: 300
Transaction Response Time: 0.25
Average Bytes per Transaction: 2500
Average Utilization for Transactions: 0.83
Number of parallel network segments: 2
Maximum LAN Through-put (bytes/second): 25000
Average LAN Through-put (bytes/second): 10000
Transactions per parallel network segment per minute: 5
Average Segment Utilization: 40 %

Document: Done

Response Time Pipe Example

- **Add to pipe:**
  - Trans workload utilization
  - Overall utilization
- **Compare:**
  - Objective
  - Estimate
  - Actual
- **Conclusions based on relationships**


RTP Section Utilizations (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop

Measurement Interval (minutes): 30
Response Times for transaction Create Account:
Objective for Overall End-to-end Response Time (seconds): 6
RTP Estimate for Overall End-to-end Response Time (seconds): 5.28
Actual Measurement of Overall End-to-end Response Time (seconds): 5.5

- ☺ The RTP estimated response time is less than the response time objective, which means it is possible for the transaction to meet the business needs. Additional anaylsis is needed to understand the effects of queuing and interference from other workloads.
- ☺ The measured response time is greater than the response time estimate, which means the estimate probably reflects the minimal transaction time and the measured time includes queuing and interference from other workloads and the RTP predictive steps can use the estimate forthe transaction service time.
- ☺ The measured response time is less than the response time objective, therefore this RTP will probably accept more transaction traffic.

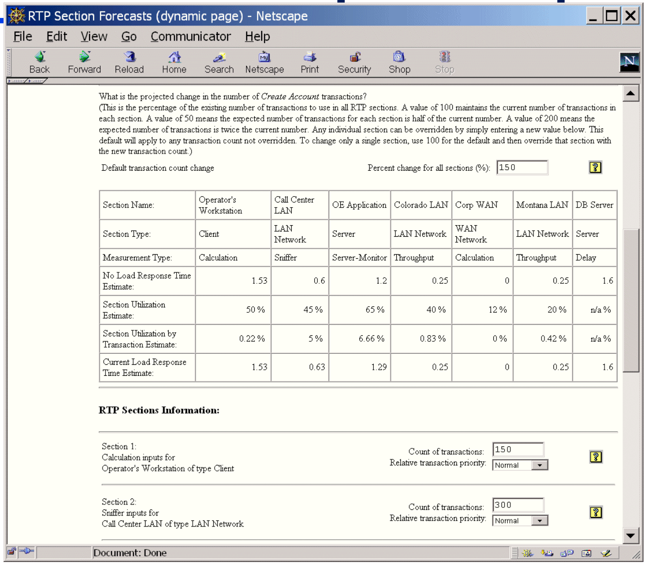
Section Name:	Rep-PC	Call Center LAN	OE Application Server	Colorado LAN	Montana LAN	DB Server
Section Type:	Client	LAN Network	Server	LAN Network	LAN Network	Server
Measurement Type:	Calculation	Sniffer	Server-Monitor	Throughput	Throughput	Delay
Response Time Estimate:	1.53	0.6	1.2	0.25	0.1	1.6
Section Utilization Estimate:	50 %	45 %	65 %	40 %	50 %	n/a %
Section Utilization by Transaction Estimate:	0.22 %	5 %	6.66 %	0.83 %	0.42 %	n/a %

Response Time Pipe Example

- **Predicting Future Response Times**
  - Use the initial response time as the service time
    - builds from the “best case” view of the transactions
    - valid because it is from very low activity time
  - Use the relative priority to control the impact of other work on transactions in the RTP section
    - only approximates the relationship
  - Use accepted queuing theory techniques
    - approximates response time (problem with high utilizations)
      - see Menascé and Allen books
    - allow override with better results (monitors, models, etc....)

Response Time Pipe Example

- **Application growth:**
  - Overall growth
  - Section growth
- **Relationship to other work in the section**
  - High
  - Normal
  - Low


RTP Section Forecasts (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

What is the projected change in the number of *Create Account* transactions?
(This is the percentage of the existing number of transactions to use in all RTP sections. A value of 100 maintains the current number of transactions in each section. A value of 50 means the expected number of transactions for each section is half of the current number. A value of 200 means the expected number of transactions is twice the current number. Any individual section can be overridden by simply entering a new value below. This default will apply to any transaction count not overridden. To change only a single section, use 100 for the default and then override that section with the new transaction count.)

Default transaction count change — Percent change for all sections (%): 150 ?

Section Name:	Operator's Workstation	Call Center LAN	OE Application	Colorado LAN	Corp WAN	Montana LAN	DB Server
Section Type:	Client	LAN Network	Server	LAN Network	WAN Network	LAN Network	Server
Measurement Type:	Calculation	Sniffer	Server-Monitor	Throughput	Calculation	Throughput	Delay
No Load Response Time Estimate:	1.53	0.6	1.2	0.25	0	0.25	1.6
Section Utilization Estimate:	50 %	45 %	65 %	40 %	12 %	20 %	n/a %
Section Utilization by Transaction Estimate:	0.22 %	5 %	6.66 %	0.83 %	0 %	0.42 %	n/a %
Current Load Response Time Estimate:	1.53	0.63	1.29	0.25	0	0.25	1.6

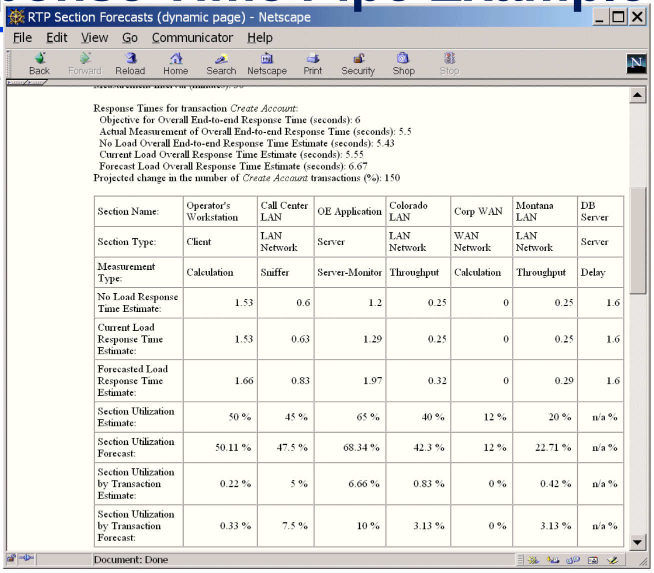
RTP Sections Information:

Section 1:
Calculation inputs for
Operator's Workstation of type Client
Count of transactions: 150
Relative transaction priority: Normal
?

Section 2:
Sniffer inputs for
Call Center LAN of type LAN Network
Count of transactions: 300
Relative transaction priority: Normal
?

Response Time Pipe Example

- **Predicting the transaction:**
  - Objective
  - Actual
  - Estimate
  - Forecast
- **Predicting each section**
  - Response
  - Utilization
  - Transaction utilization


RTP Section Forecasts (dynamic page) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Document: Done

Response Times for transaction *Create Account*:
Objective for Overall End-to-end Response Time (seconds): 6
Actual Measurement of Overall End-to-end Response Time (seconds): 5.5
No Load Overall End-to-end Response Time Estimate (seconds): 5.43
Current Load Overall Response Time Estimate (seconds): 5.55
Forecast Load Overall Response Time Estimate (seconds): 6.67
Projected change in the number of *Create Account* transactions (%): 150

Section Name:	Operator's Workstation	Call Center LAN	OE Application	Colorado LAN	Corp WAN	Montana LAN	DB Server
Section Type:	Client	LAN Network	Server	LAN Network	WAN Network	LAN Network	Server
Measurement Type:	Calculation	Sniffer	Server-Monitor	Throughput	Calculation	Throughput	Delay
No Load Response Time Estimate:	1.53	0.6	1.2	0.25	0	0.25	1.6
Current Load Response Time Estimate:	1.53	0.63	1.29	0.25	0	0.25	1.6
Forecasted Load Response Time Estimate:	1.66	0.83	1.97	0.32	0	0.29	1.6
Section Utilization Estimate:	50 %	45 %	65 %	40 %	12 %	20 %	n/a %
Section Utilization Forecast:	50.11 %	47.5 %	68.34 %	42.3 %	12 %	22.71 %	n/a %
Section Utilization by Transaction Estimate:	0.22 %	5 %	6.66 %	0.83 %	0 %	0.42 %	n/a %
Section Utilization by Transaction Forecast:	0.33 %	7.5 %	10 %	3.13 %	0 %	3.13 %	n/a %

Questions?

- **References:**
  - Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning

    Daniel A. Menascé, Virgilio A. F. Almeida.

    Prentice Hall, 2000. ISBN: 0130863289
  - Probability, Statistics and Queueing Theory With Computer Science Applications

    Allen, Arnold O.

    Academic Press, 1990. ISBN: 0120510510

**End-To-End Scaling and The Response Time Pipe are service marks of Simalytic Solutions, LLC.
All other trademarked names and terms are the property of their respective owners.**