

End-To-End Scaling: The Response Time Pipe

Rocky Mountain CMG, March 8, 2002 http://www.simalytic.com/RMCMG02ppt.pdf

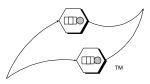
Dr. Tim R. Norton
Simalytic Solutions, LLC

719-635-5825

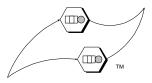
email: tim.norton@simalytic.com http://www.simalytic.com



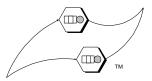
- What's the Problem
 - Background
- The Response Time Pipe Solution
 - Techniques that fit the problem
- Response Time Pipe Example
 - Sample solution to a hypothetical situation



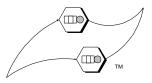
- How does the performance of a computer application effect the business?
 - Defining the relationship between the two:
 - The business result when the application changes
 - The application result when the business changes
 - What is the "effect"?
 - Requires measuring both
 - Implies there is a "good" 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?



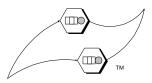
- Measure the "effect"
 - Measure the Pieces
 - Measuring the application
 - Different types of applications
 - Fat/thin client, multi-tier, web based, proprietary, ...
 - Different units of work
 - ▲ Transactions, messages, interactive, asynchronous, ...
 - What is the end-user's experience?
 - Measure everything or just what's "important"?



- 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

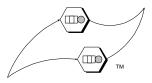


- Measure the "effect"
 - Measure the Business
 - Measuring the response time
 - Component response times lack continuity
 - ▲ Pitfall: viewing the magnitude of the component change as the magnitude of the business change
 - End-to-end response times lack enough detail
 - Hard to correlate ETE-RT across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



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



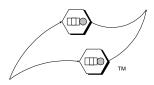
Predicting when the effect will be "bad"

- 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 Needed in a Solution?

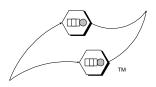
- Need an approximation technique
 - Easy to use without years of experience
 - Identifies areas of concern
 - Eliminates areas that don't matter (right now)
 - Usable results quickly enough for business decisions
- Need a technique to tie all the measurement pieces together, regardless of sources
- 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?

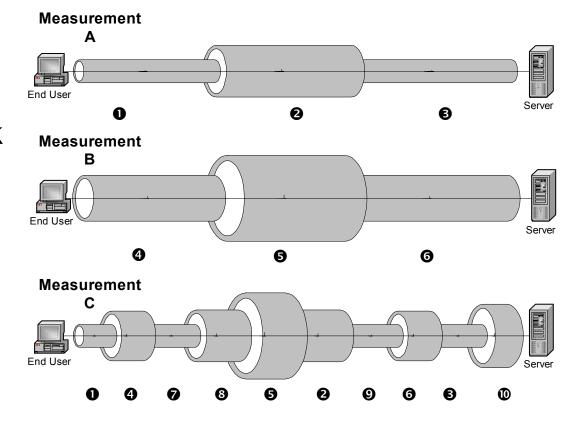
- Way to visualize the relationships between components used by an application.
- A technique that quickly connects different types of component performance measurements or approximations.
- A technique to relate the performance of the components to the business objective.



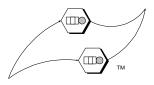
Response Time Pipe Solution

Why a Pipe?

- To provide a visual framework that expresses:
 - Capacity
 - Connection
 - Flow
 - Sections
 - Constrictions



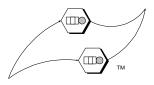
 Looking at different sections provides different perceptions of capacity and performance



Response Time Pipe Solution

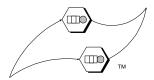
How to Build an RTP

- 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

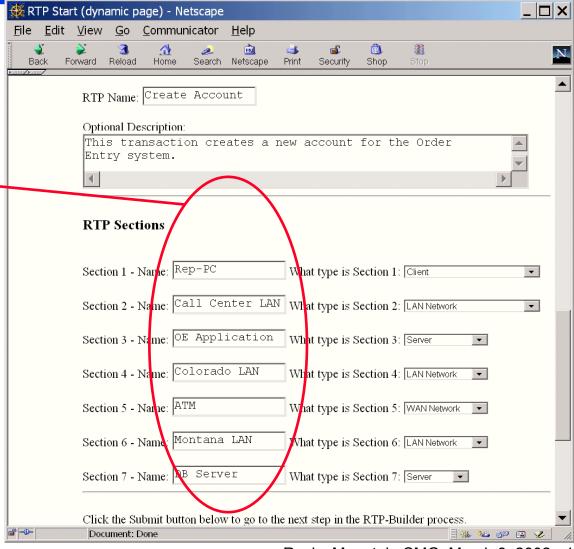


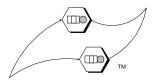
Hypothetical Situation and Infrastructure

- Operators service customers in a call center
- Simple <u>Create Account</u> Transaction
- Multi-tier infrastructure
 - Client PC
 - Call Center LAN
 - Order Entry Application Server
 - Network segments (LAN→WAN→LAN)
 - Database Sever

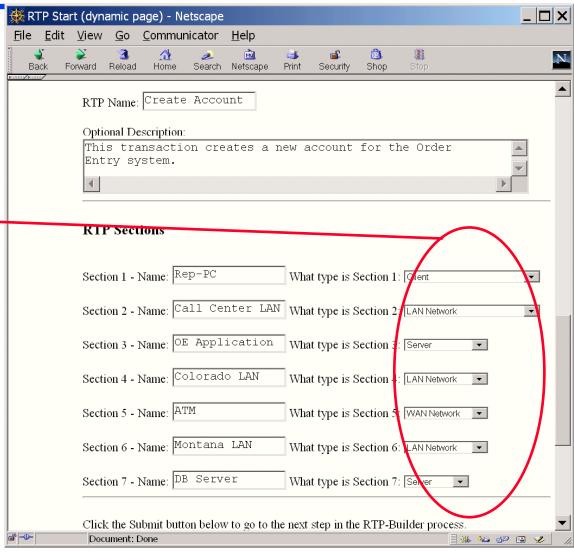


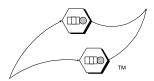
- Define each section of the RTP
 - Name-
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



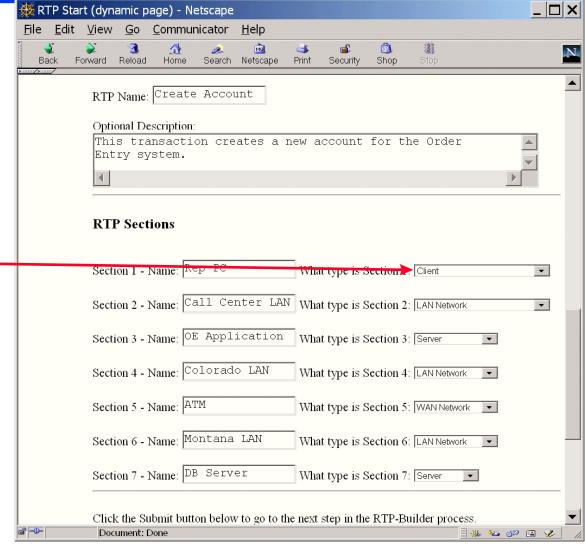


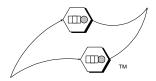
- Define each section of the RTP
 - Name
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



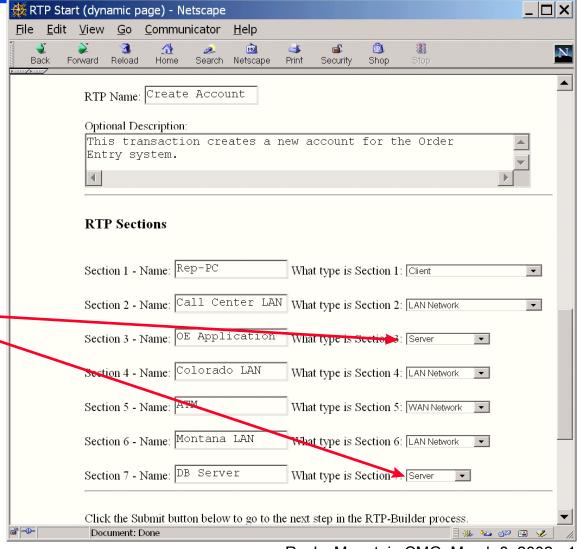


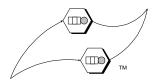
- Define each section of the RTP
 - Name
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



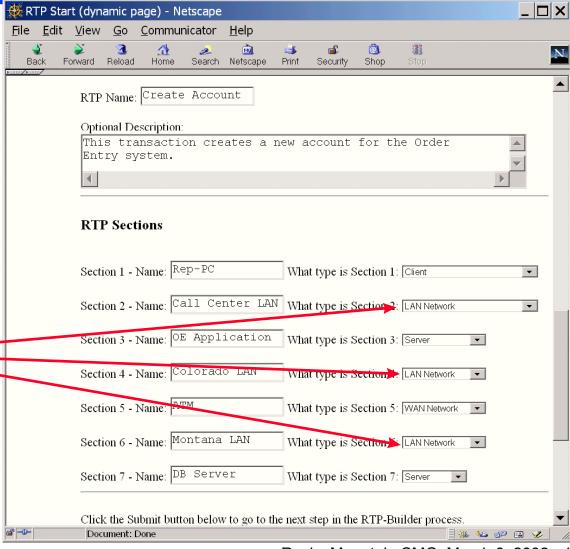


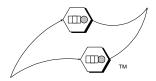
- Define each section of the RTP
 - Name
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



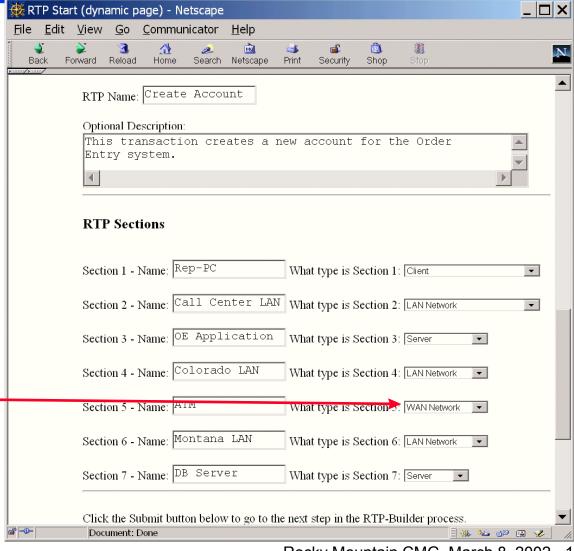


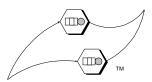
- Define each section of the RTP
 - Name
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



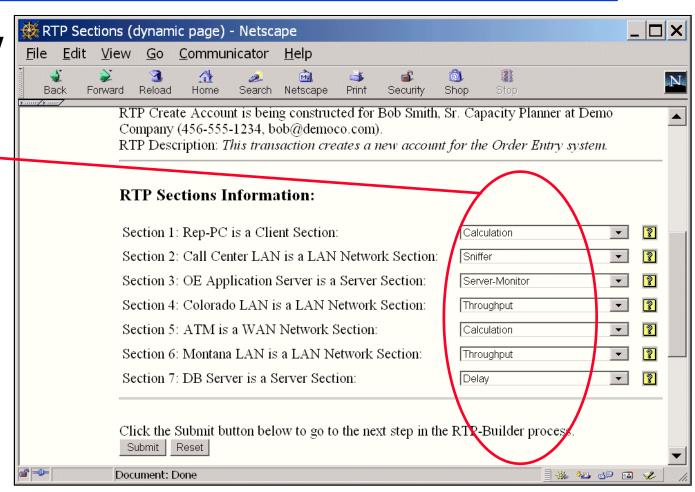


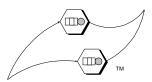
- Define each section of the RTP
 - Name
 - Type of section
 - Client
 - Server
 - LAN
 - WAN



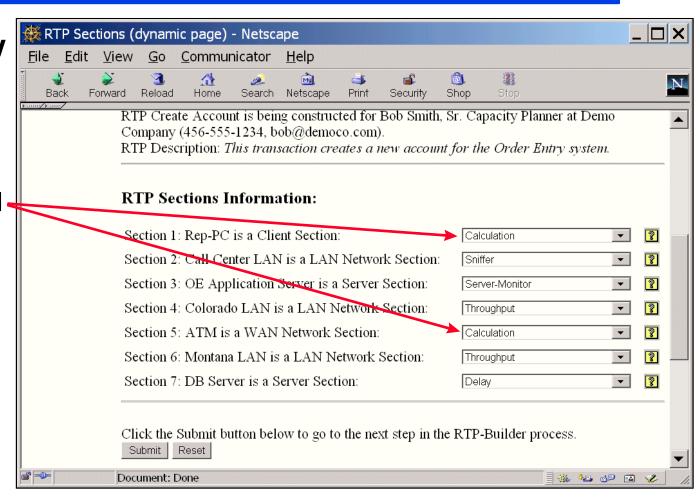


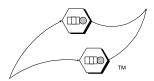
- Define how each section is measured
 - Calculated
 - Sniffer
 - Monitor
 - Throughput
 - Delay





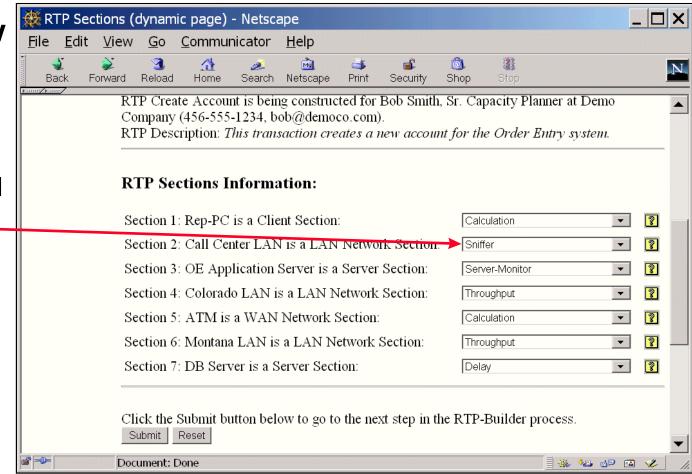
- Define how each section is measured
 - Calculated
 - Sniffer
 - Monitor
 - Throughput
 - Delay

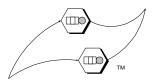




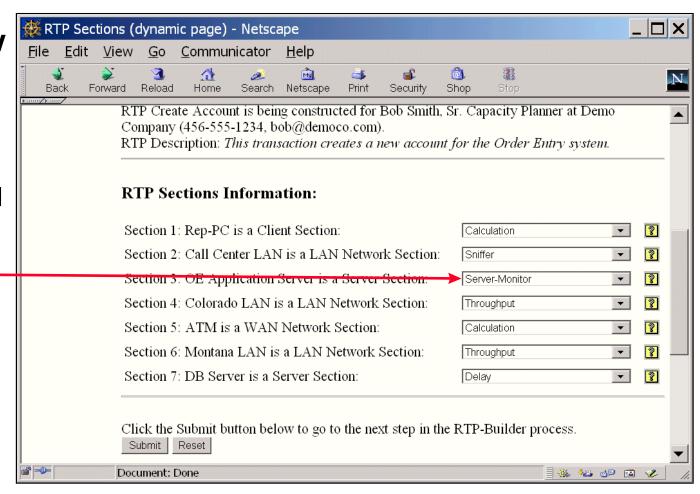
Define how each section is measured

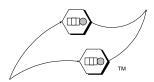
- Calculated
- Sniffer -
- Monitor
- Throughput
- Delay





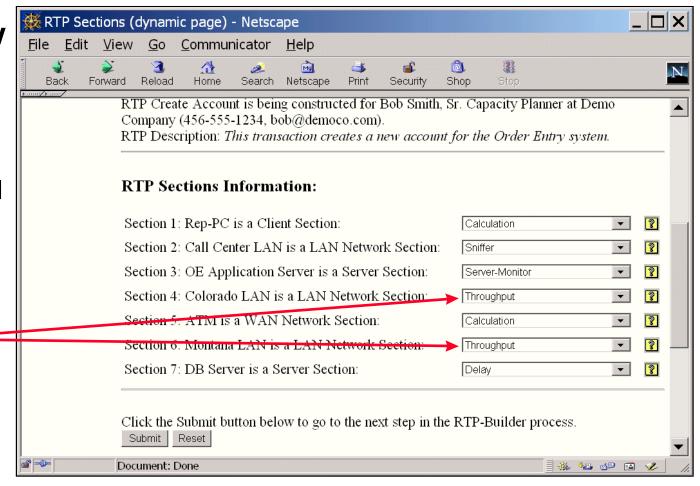
- Define how each section is measured
 - Calculated
 - Sniffer
 - Monitor-
 - Throughput
 - Delay

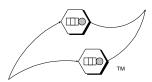




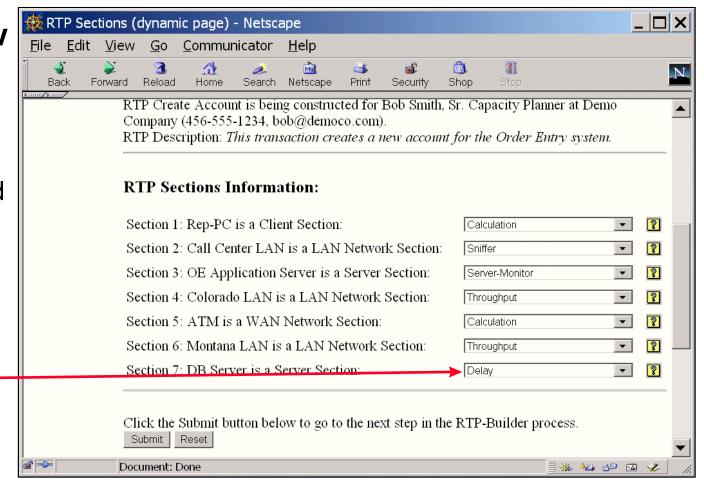
Define how each section is measured

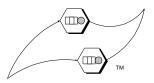
- Calculated
- Sniffer
- Monitor
- Throughput ——
- Delay



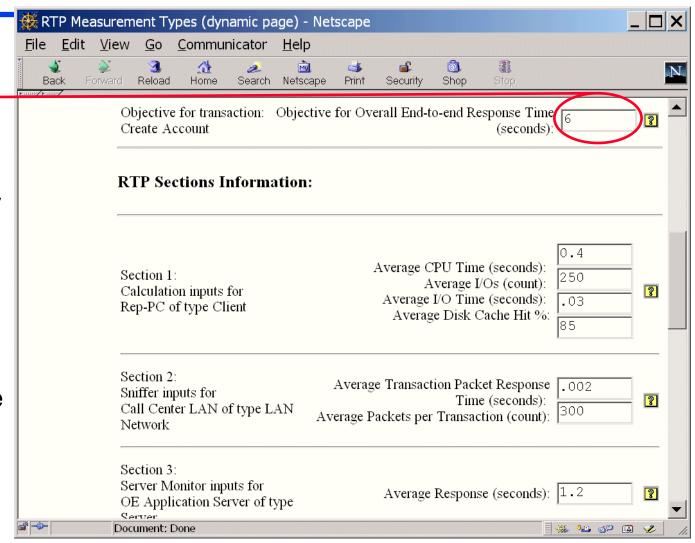


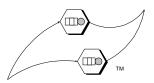
- Define how each section is measured
 - Calculated
 - Sniffer
 - Monitor
 - Throughput
 - Delay-



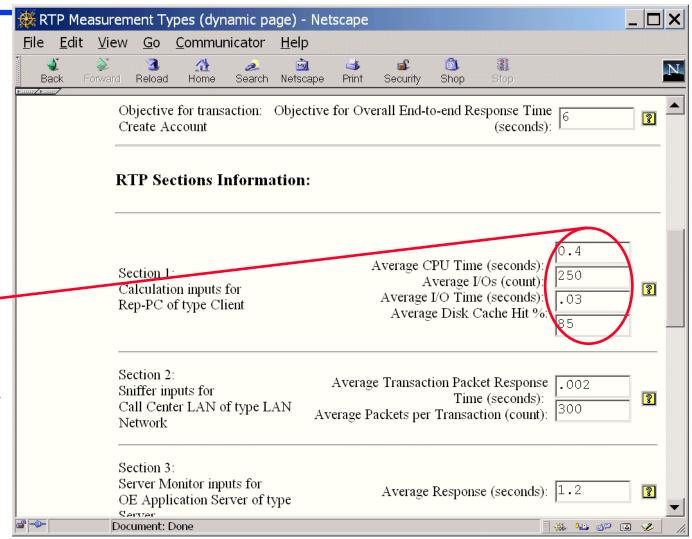


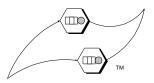
- Overall objective
- Enter the transaction measures for each section
 - Client calc: CPU & I/O
 - Sniffer:
 Packet count and response time
 - Monitor: measured value



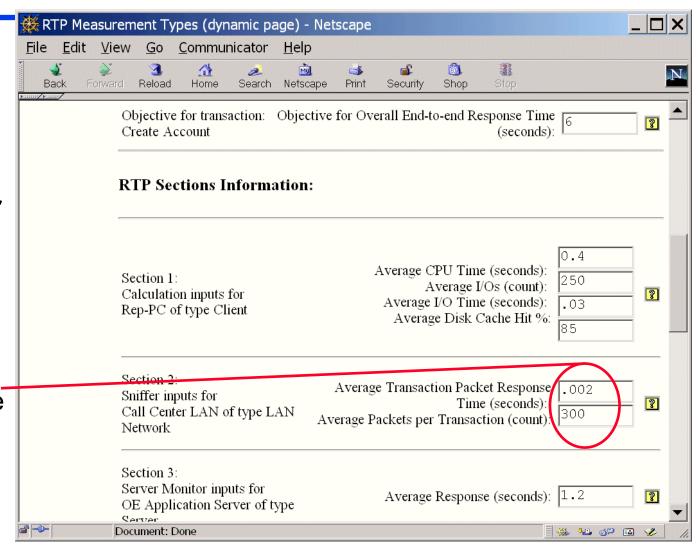


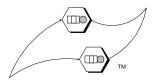
- Overall objective
- Enter the transaction measures for each section
 - Client calc: CPU & I/O
 - Sniffer:
 Packet count and response time
 - Monitor: measured value



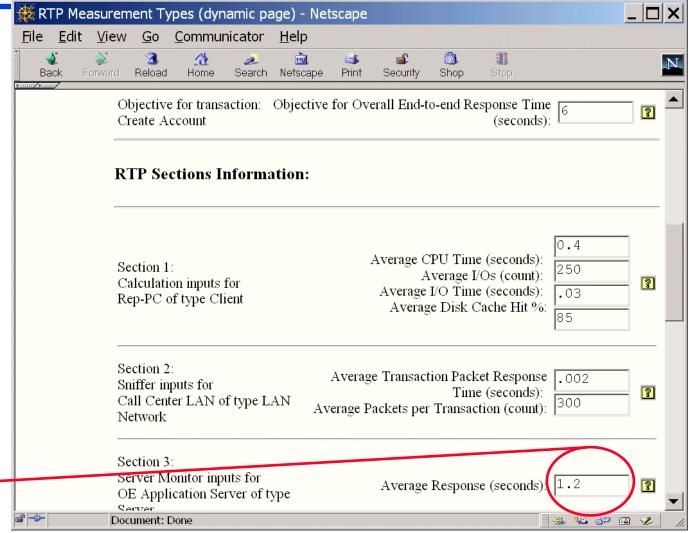


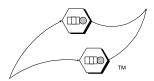
- Overall objective
- Enter the transaction measures for each section
 - Client calc: CPU & I/O
 - Sniffer:
 Packet count_and response time
 - Monitor: measured value



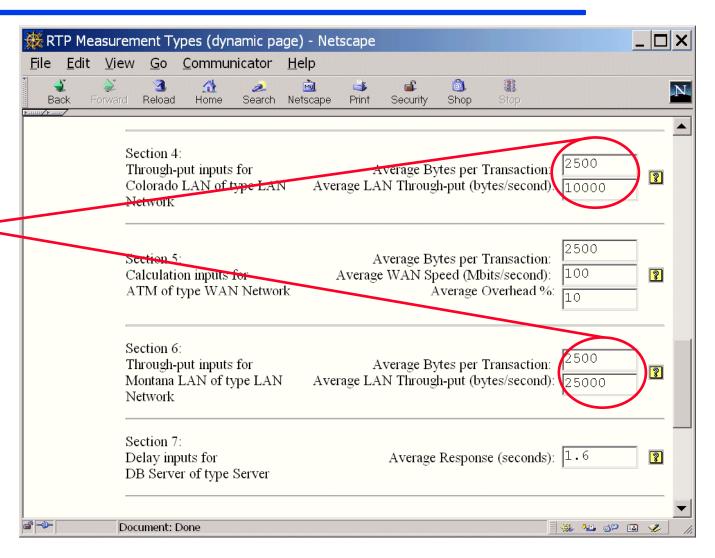


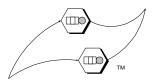
- Overall objective
- Enter the transaction measures for each section
 - Client calc: CPU & I/O
 - Sniffer:
 Packet count and response time
 - Monitor: measured value



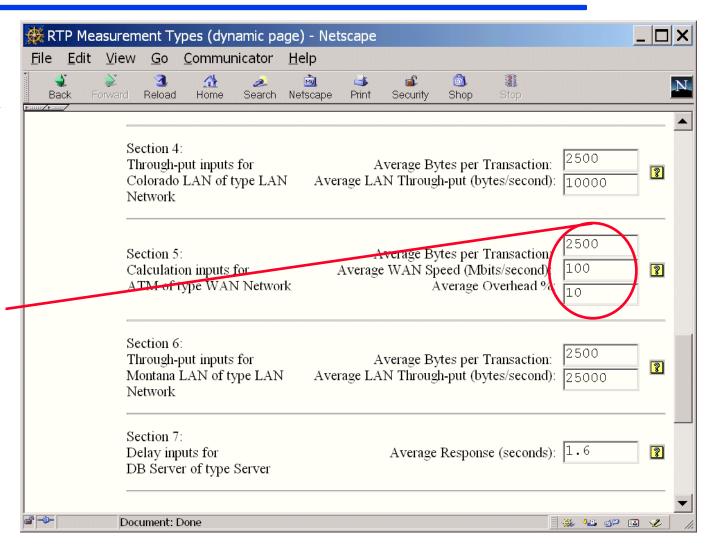


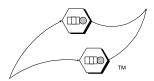
- Enter the transaction measures for each section
 - Through-put: bytes and through-put
 - WAN calc: bytes, speed and overhead
 - Delay: value



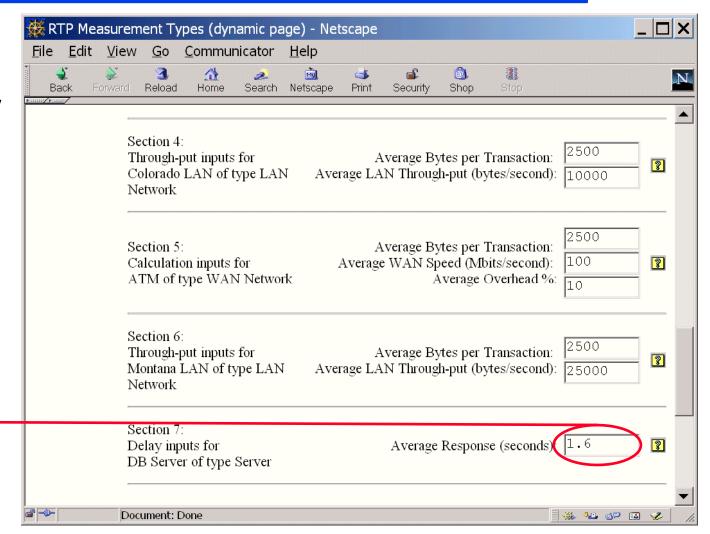


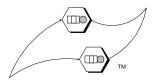
- Enter the transaction measures for each section
 - Through-put: bytes and through-put
 - WAN calc: bytes, speed and overhead
 - Delay: value



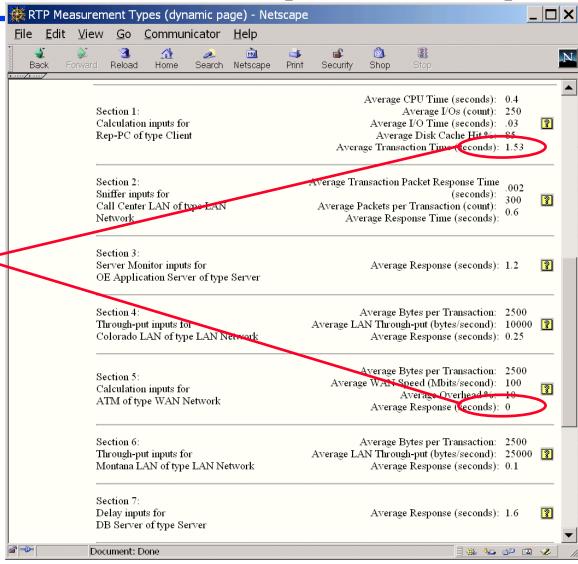


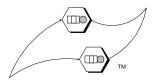
- Enter the transaction measures for each section
 - Through-put: bytes and through-put
 - WAN calc: bytes, speed and overhead
 - Delay:value -



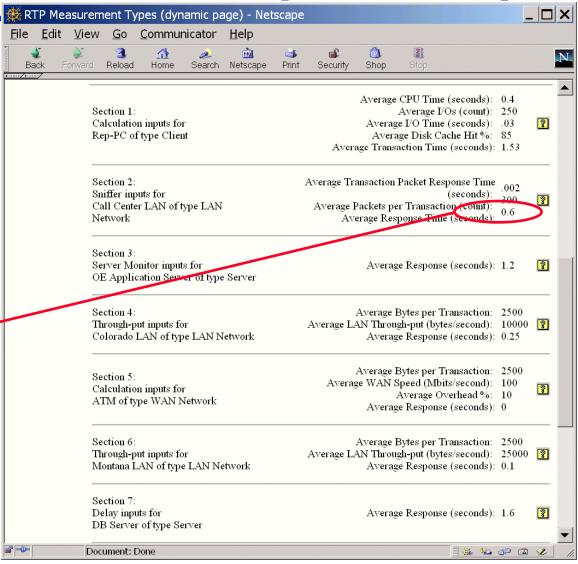


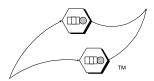
- Calculate the transaction response times for each section
 - Calc: add the component times
 - Sniffer: packet response time * count
 - Monitor: value
 - Through-put: based on total bytes
 - Delay: value



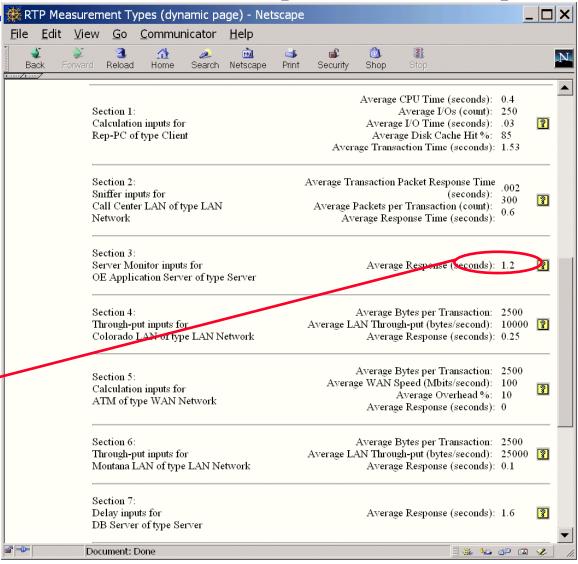


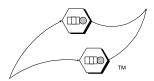
- Calculate the transaction response times for each section
 - Calc: add the component times
 - Sniffer: packet responsetime * count
 - Monitor: value
 - Through-put: based on total bytes
 - Delay: value



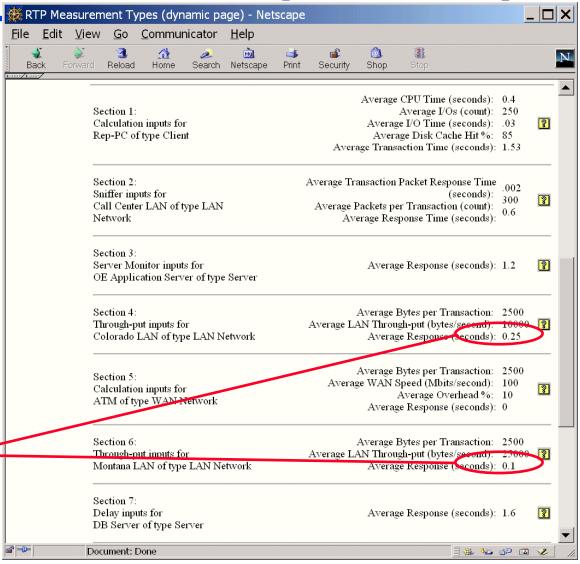


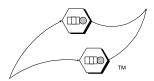
- Calculate the transaction response times for each section
 - Calc: add the component times
 - Sniffer: packet response time * count
 - Monitor: value
 - Through-put: based on total bytes
 - Delay: value



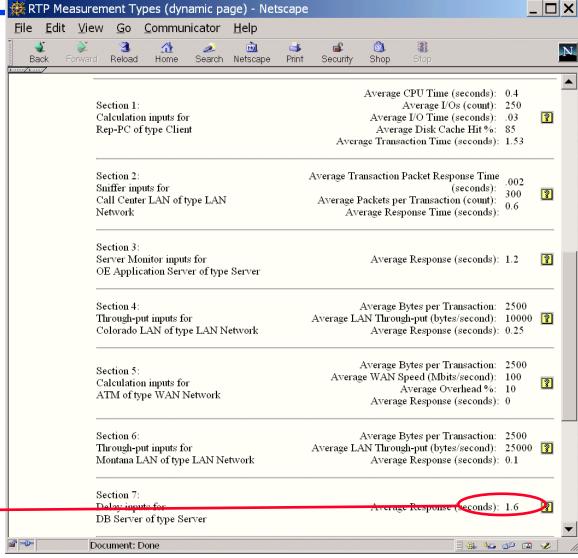


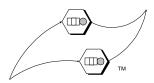
- Calculate the transaction response times for each section
 - Calc: add the component times
 - Sniffer: packet response time * count
 - Monitor: value
 - Through-put: based on total bytes
 - Delay: value



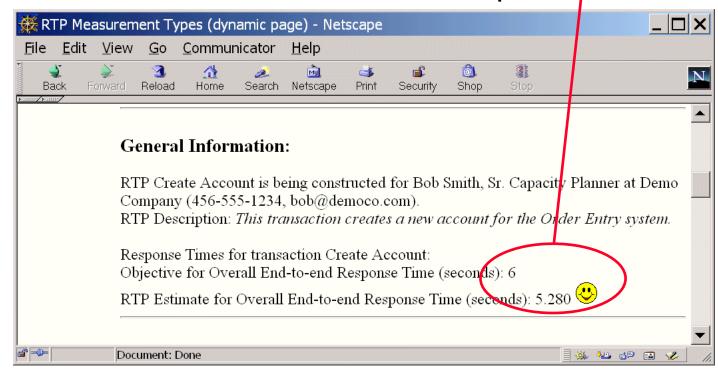


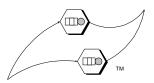
- Calculate the transaction response times for each section
 - Calc: add the component times
 - Sniffer: packet response time * count
 - Monitor: value
 - Through-put: based on total bytes
 - Delay: value



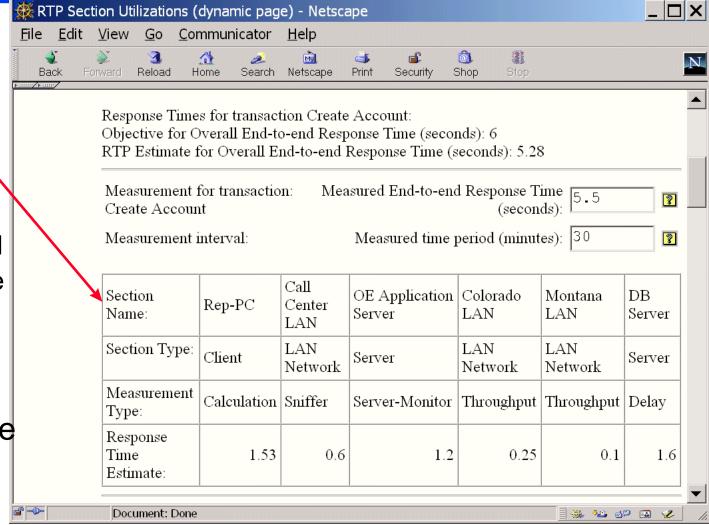


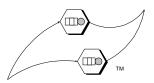
- Compare the estimate to the objective
 - First indicator of "goodness" or "badness"
 - "Best case" estimate of transaction response time



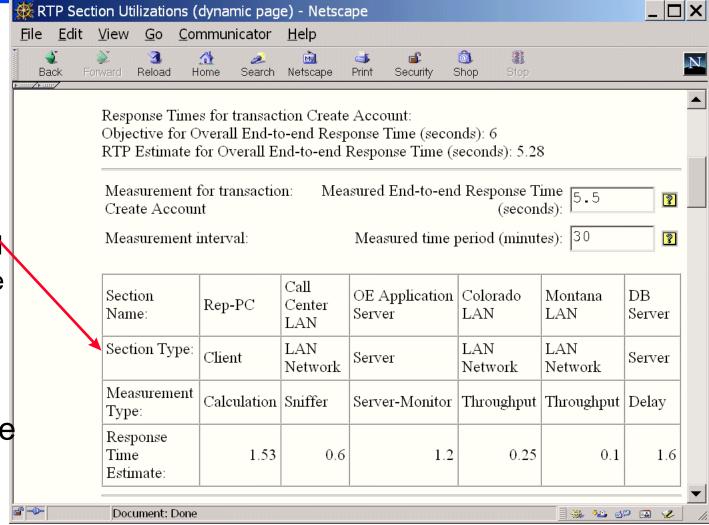


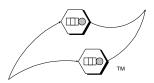
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval



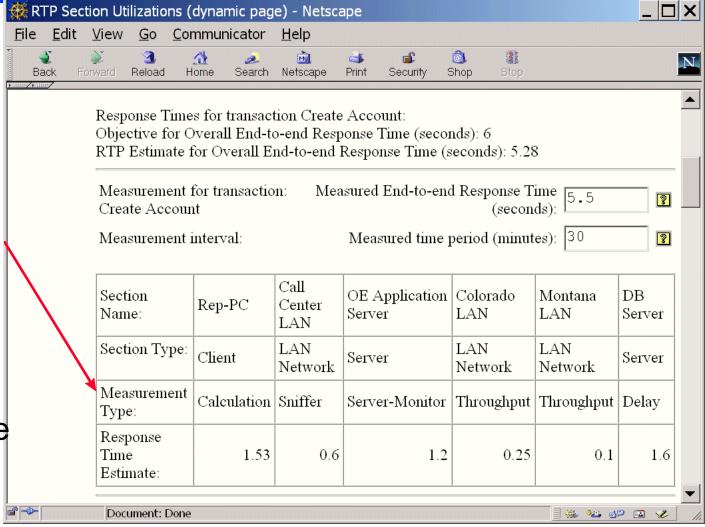


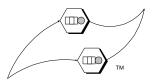
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval



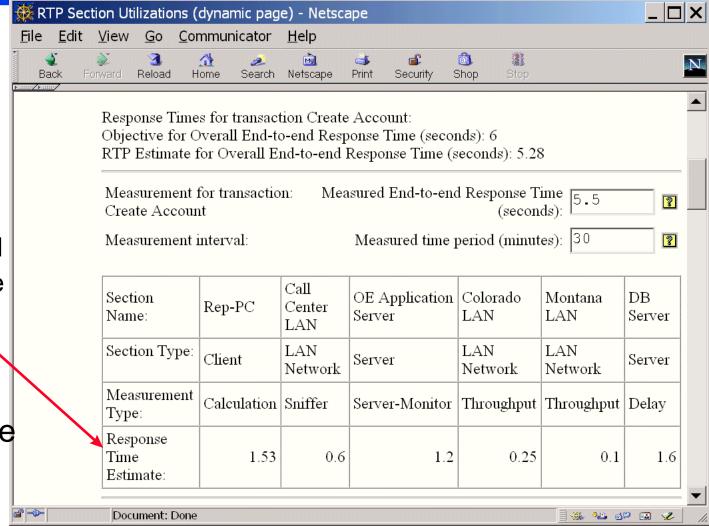


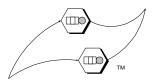
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval



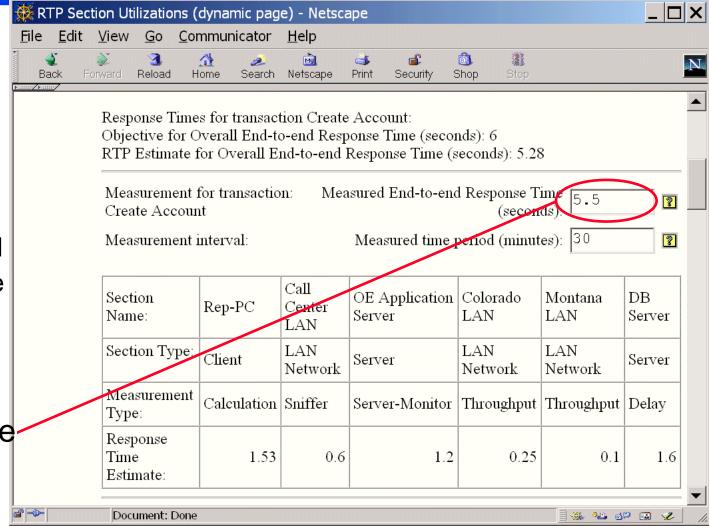


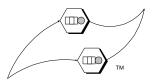
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval



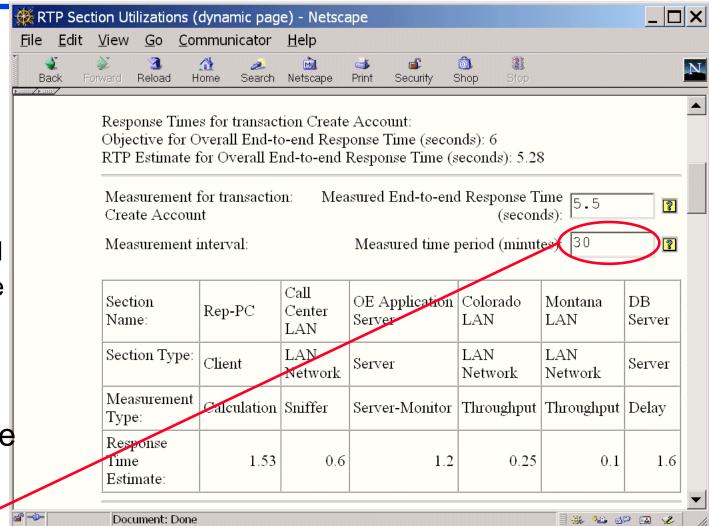


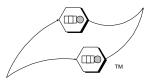
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval



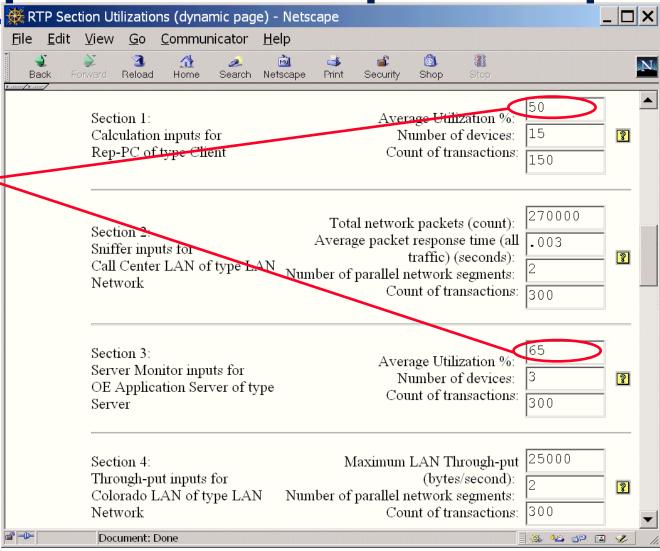


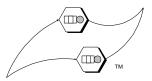
- Build the pipe
 - Each section
 - Type
 - How it's measured
 - Response times
- Measured:
 - Overall response time
 - Interval-



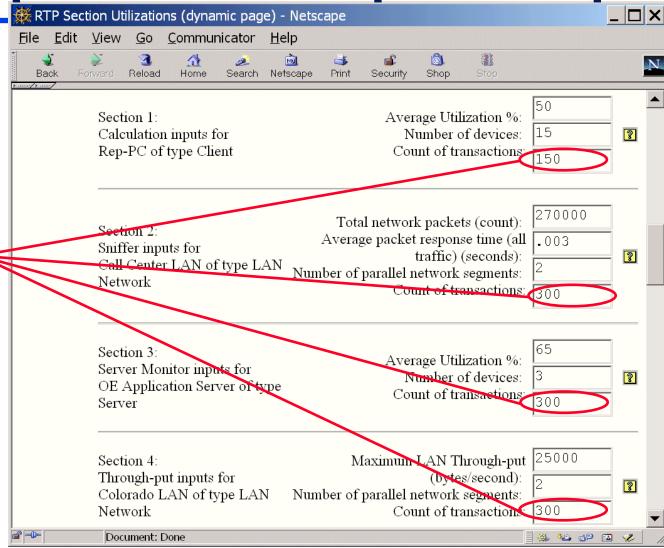


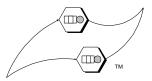
- Add current load information
 - utilizations
 - transaction counts
 - packet counts
 - byte counts
 - parallelism



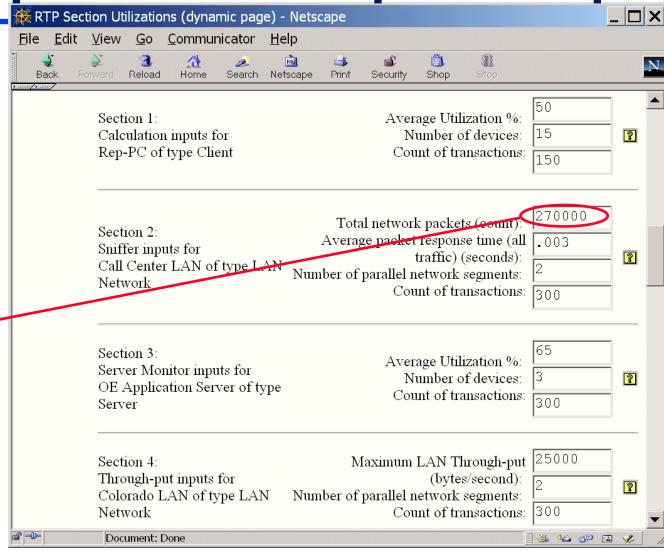


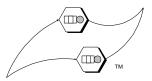
- Add current load information
 - utilizations
 - transaction counts
 - packet counts
 - byte counts
 - parallelism



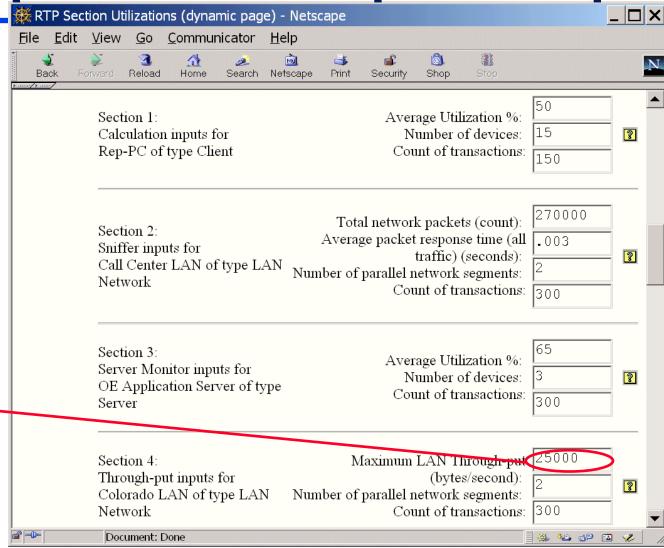


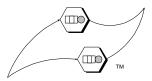
- Add current load information
 - utilizations
 - transaction counts
 - packet counts
 - byte counts
 - parallelism



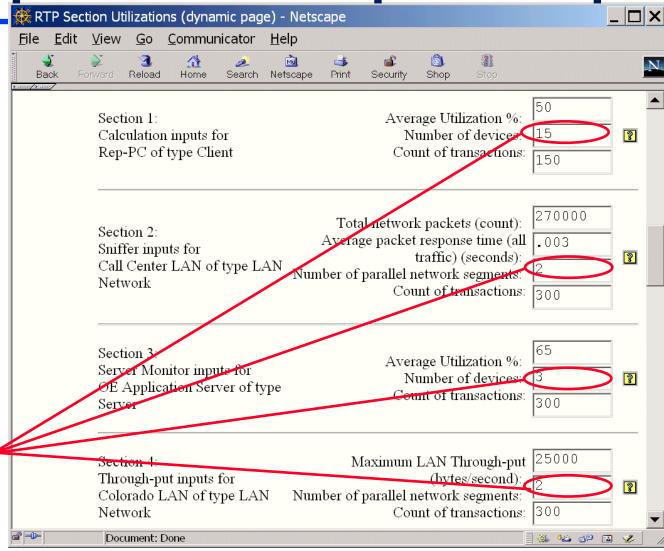


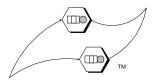
- Add current load information
 - utilizations
 - transaction counts
 - packet counts
 - byte counts
 - parallelism





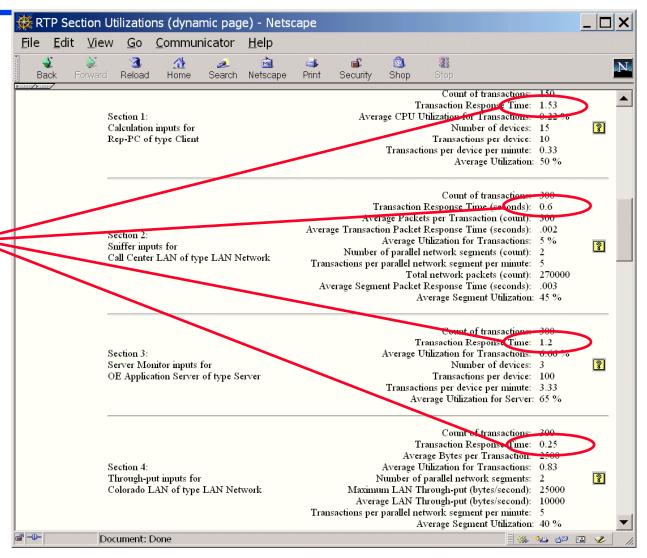
- Add current load information
 - utilizations
 - transaction counts
 - packet counts
 - byte counts
 - parallelism

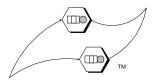




Calculations for each section

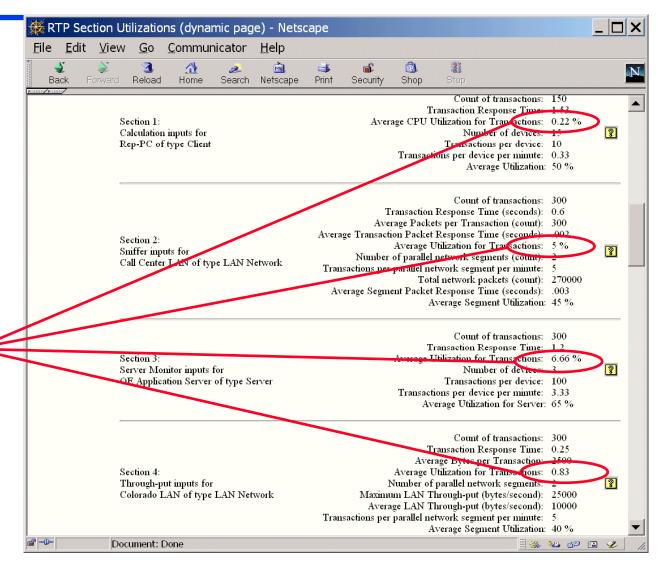
- New transaction response times
- Transaction workload utilization
- Overall utilization
- Accounts for effect of current load

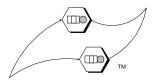




Calculations for each section

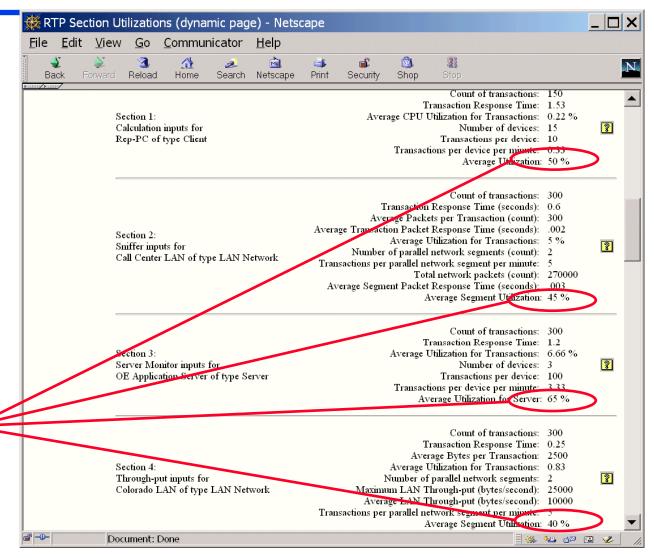
- New transaction response times
- Transaction workload utilization
- Overall utilization
- Accounts for effect of current load

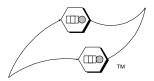




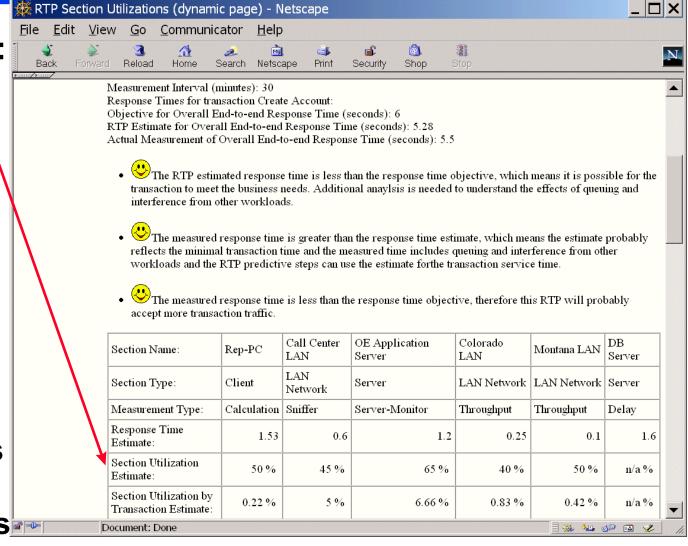
Calculations for each section

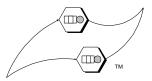
- New transaction response times
- Transaction workload utilization
- Overall utilization
- Accounts for effect of current load



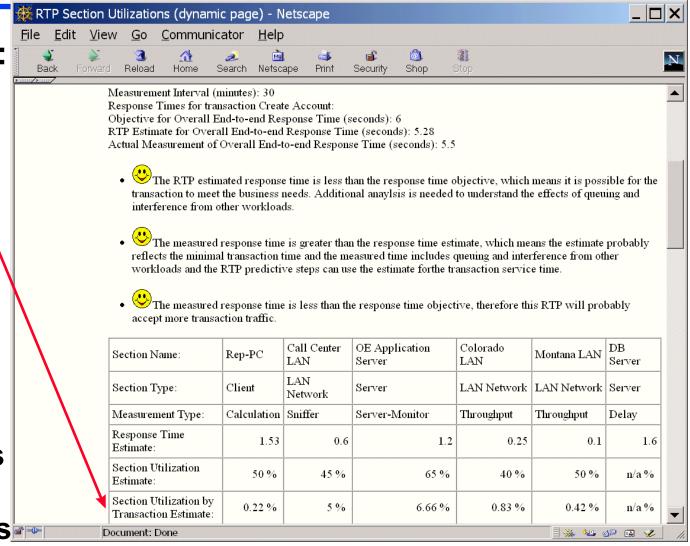


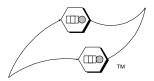
- Add to pipe:
 - Trans workload utilization
 - Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions based on relationships





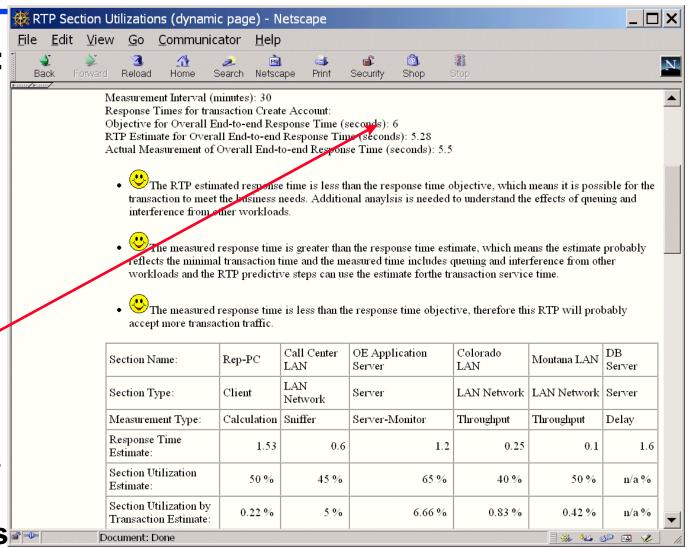
- Add to pipe:
 - Trans workload utilization
 - Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions based on relationships

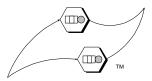




Add to pipe:

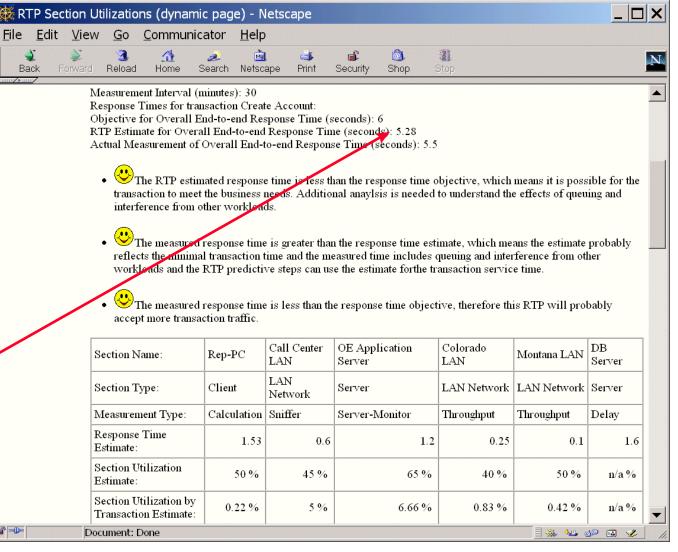
- Trans workload utilization
- Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions based on relationships

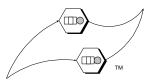




Add to pipe:

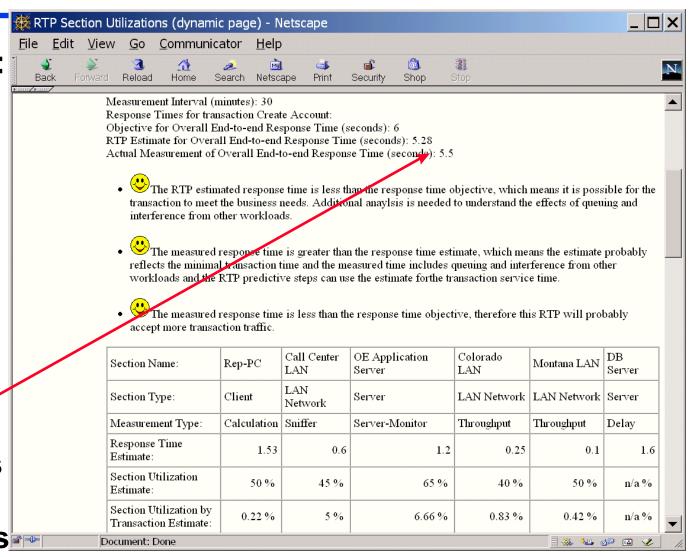
- Trans workload utilization
- Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions based on relationships

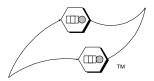




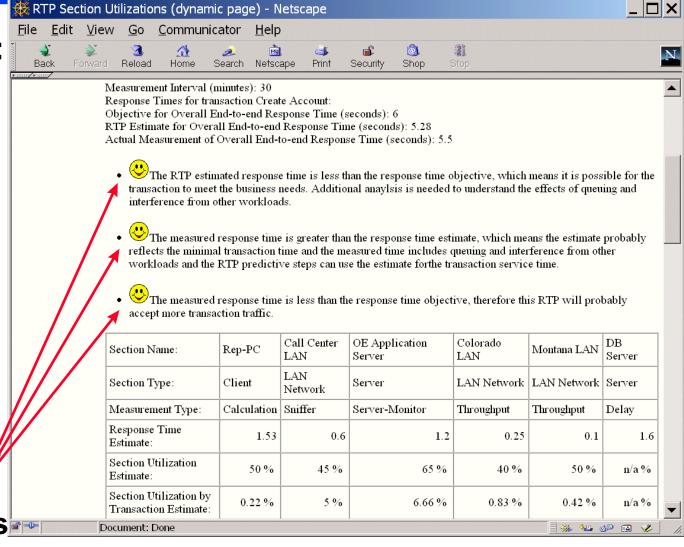
Add to pipe:

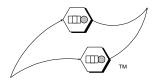
- Trans workload utilization
- Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions based on relationships





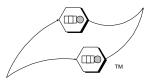
- Add to pipe:
 - Trans workload utilization
 - Overall utilization
- Compare:
 - Objective
 - Estimate
 - Actual
- Conclusions
 based on
 relationships



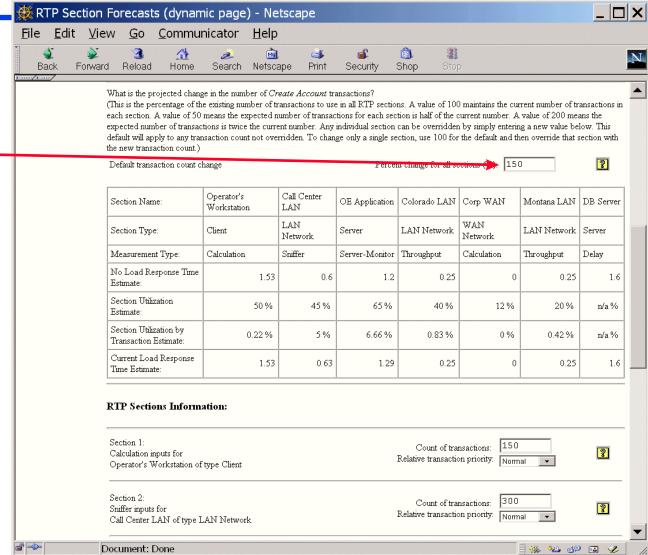


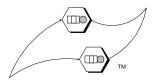
Predicting Future Response Times

- 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....)

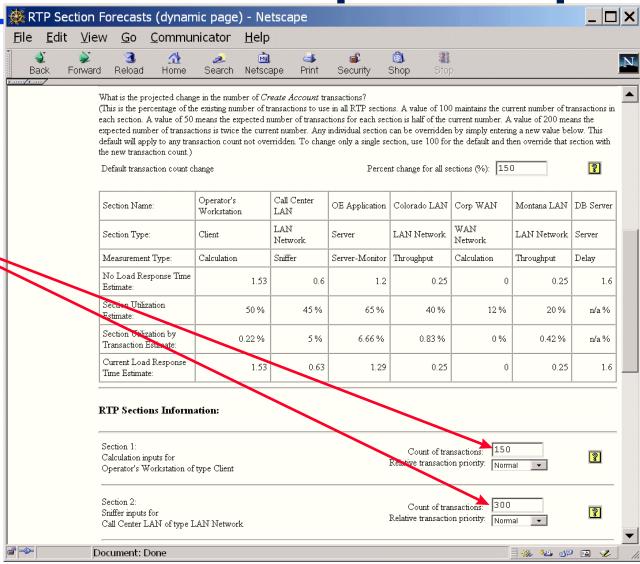


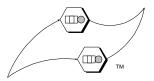
- Application growth:
 - Overall growth
 - Section growth
- Relationship to other work in the section
 - High
 - Normal
 - Low



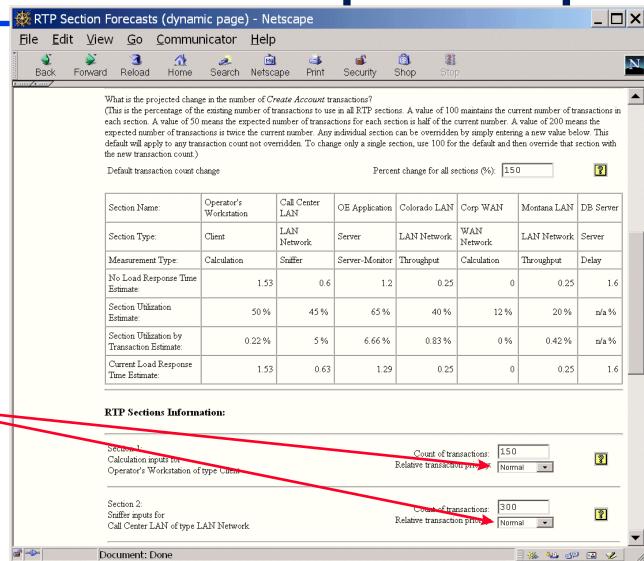


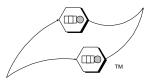
- Application growth:
 - Overall growth
 - Section growth
- Relationship to other work in the section
 - High
 - Normal
 - Low



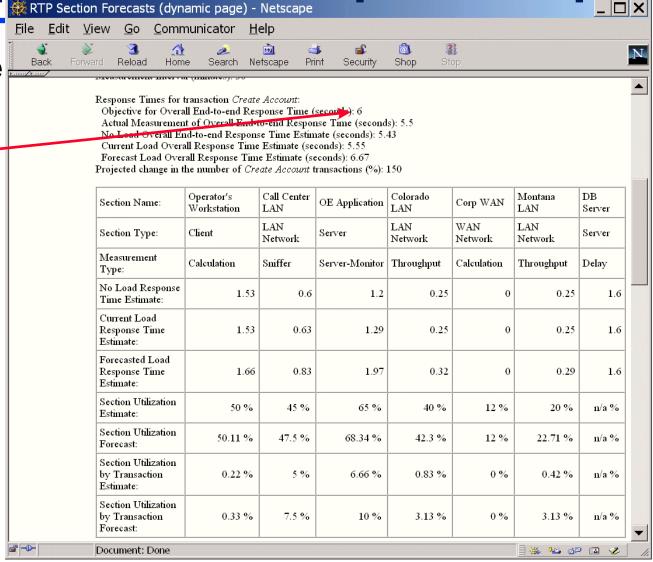


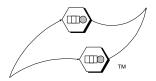
- Application growth:
 - Overall growth
 - Section growth
- Relationship to other work in the section
 - High
 - Normal
 - Low



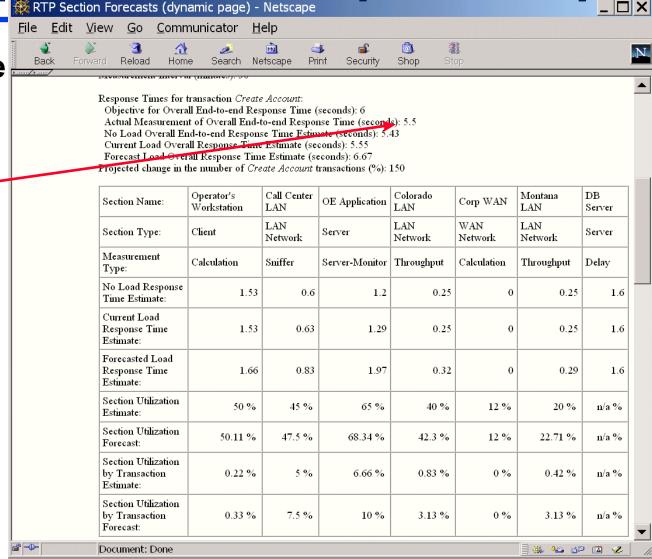


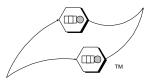
- Objective
- Actual
- Estimate
- Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization



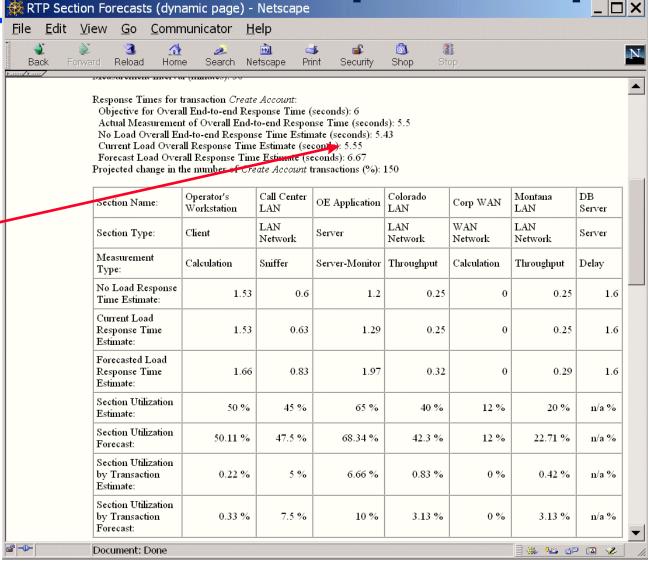


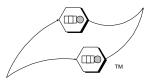
- Objective
- Actual -
- Estimate
- Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization



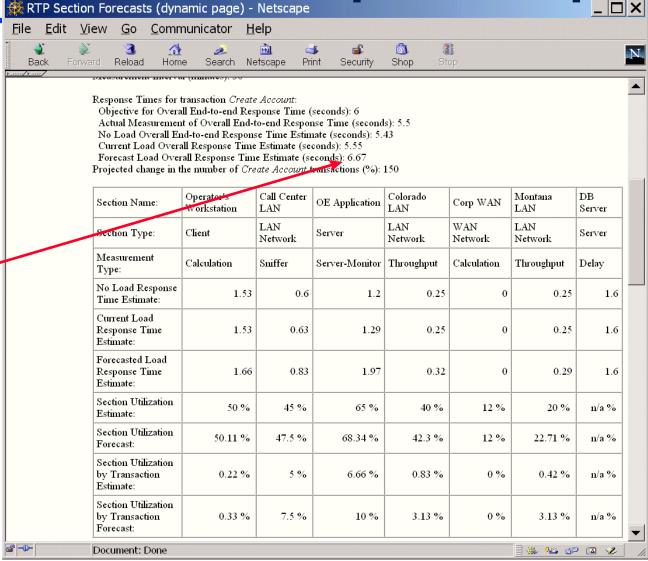


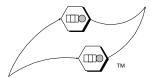
- Objective
- Actual
- Estimate
- Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization



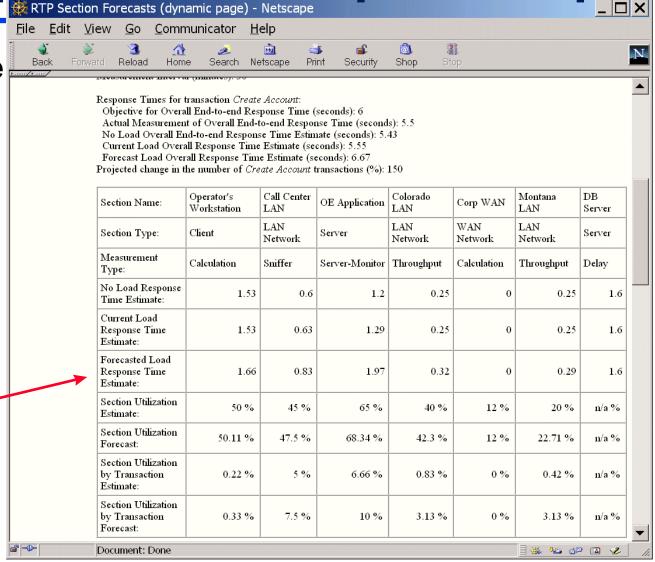


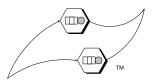
- Objective
- Actual
- Estimate
- Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization



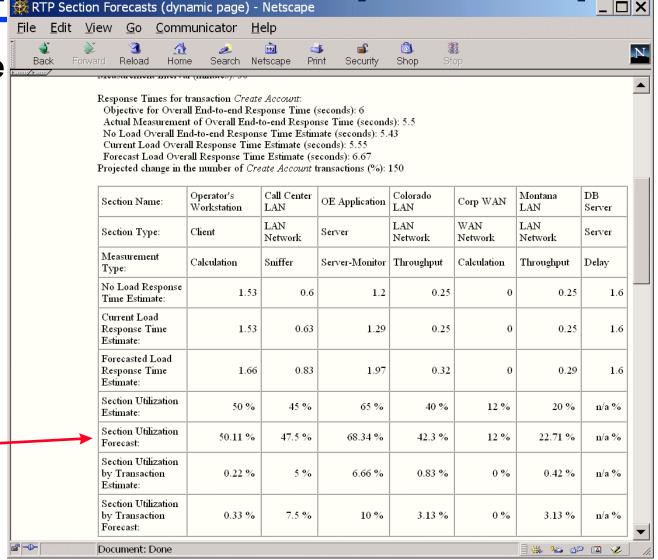


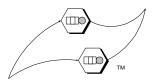
- Predicting the transaction:
 - Objective
 - Actual
 - Estimate
 - Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization



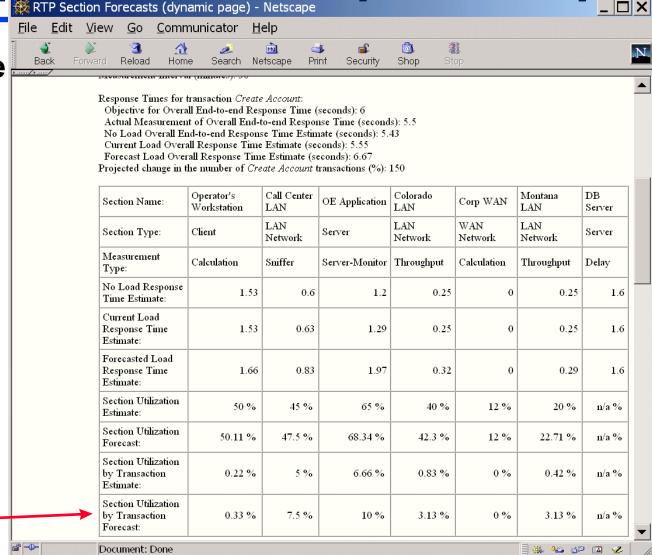


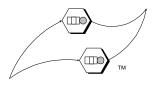
- Predicting the transaction:
 - Objective
 - Actual
 - Estimate
 - Forecast
- Predicting each section
 - Response
 - Utilization
 - Transaction utilization





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





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.