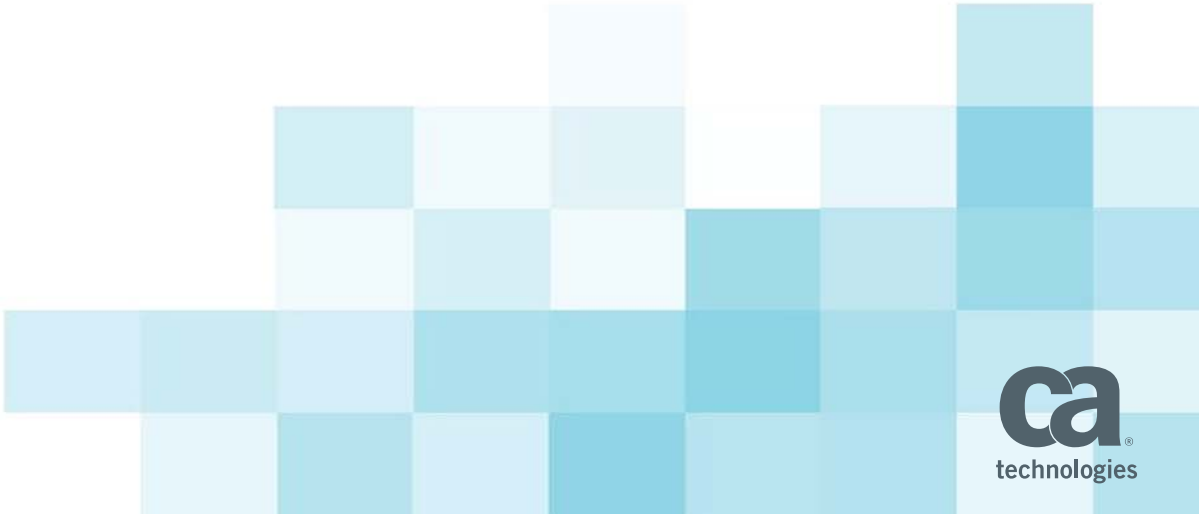
EM DevXchange 11 & 12th May 2016
ADA/MTP Best Practices
Infrastructure Management

Todor Kardjiev
Principal Consultant Technical Sales - EMEA

14 May 2016



Disclaimer

Certain information in this presentation may outline CA's general product direction. This presentation shall not serve to (i) affect the rights and/or obligations of CA or its licensees under any existing or future license agreement or services agreement relating to any CA software product; or (ii) amend any product documentation or specifications for any CA software product. This presentation is based on current information and resource allocations as of January 5, 2016 and **is subject to change or withdrawal by CA at any time without notice**. **The development, release and timing of any features or functionality described in this presentation remain at CA's sole discretion**.

Notwithstanding anything in this presentation to the contrary, upon the general availability of any future CA product release referenced in this presentation, CA may make such release available to new licensees in the form of a regularly scheduled major product release. Such release may be made available to licensees of the product who are active subscribers to CA maintenance and support, on a when and if-available basis. The information in this presentation is not deemed to be incorporated into any contract.

THIS PRESENTATION IS FOR YOUR INFORMATIONAL PURPOSES ONLY. CA assumes no responsibility for the accuracy or completeness of the information. TO THE EXTENT PERMITTED BY APPLICABLE LAW, CA PROVIDES THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. **In n**o event will CA be liable for any loss or damage, direct or indirect, in connection with this presentation, including, without limitation, lost profits, lost investment, business interruption, goodwill, or lost data, even if CA is expressly advised in advance of the possibility of such damages.

Agenda

1. ADA/MTP BACK TO BASICS
2. ARCHITECTURE
3. PLACEMENT – WHERE AND WHY
4. CONFIGURATION
5. DATA ANALYSIS
6. Q & A

ADA/MTP Back to basics

ADA – What it is, and What it is Not

- What it is --
  - Application performance from the network's perspective

"How well is my application infrastructure delivering application performance?"

- What it is not –
  - NO Business-transaction "page load" response times
  - NO visibility inside the application

What problems does it solve?

Deliver consistent application performance

Mitigate the risk of change

Solve performance problems faster

Avoid & reduce infrastructure costs

Make more informed infrastructure decisions

Prove network, server & application performance

ca technologies

How does ADA compliment our APM solution set?

- Passive application performance from the network's perspective
- Monitor any TCP-based application
  - Java & .NET, Oracle, Citrix, Other
- Isolate performance issue to proper domain
- ADA Multiport collector (MTP) provides multiple collection points & 10 Gig capability for Wily CEM TIM functionality.
- Data & navigational integration with Introscope & CEM
- APM + ADA = 360 degree visibility from the end-user experience, through the application, across the network, right down to the packets involved!

How does it work?

- ADA passively analyzes TCP Header information
- ADA uses “network + server + TCP-port” triplet combinations to baseline performance
- Performance is broken out by Network, Server, Application
  - Network = client subnets
  - Server = server that is serving the content
  - Application = which TCP port(s) does the server listen on

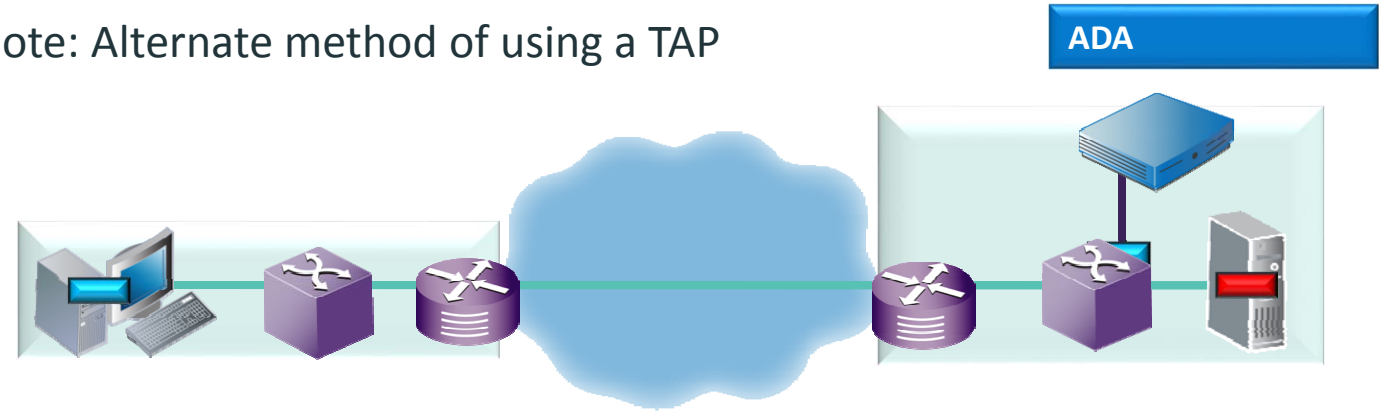
How does it work?

- Performance analysis works for both client-to-data center analysis as well as nTier application environment analysis
  - Network = front-end server
  - Server = app server (next tier of app)
  - Application = which TCP port(s) the app server listens on

How does it work?

ADA passively monitors TCP traffic mirrored to it by commands configured within an Ethernet switch

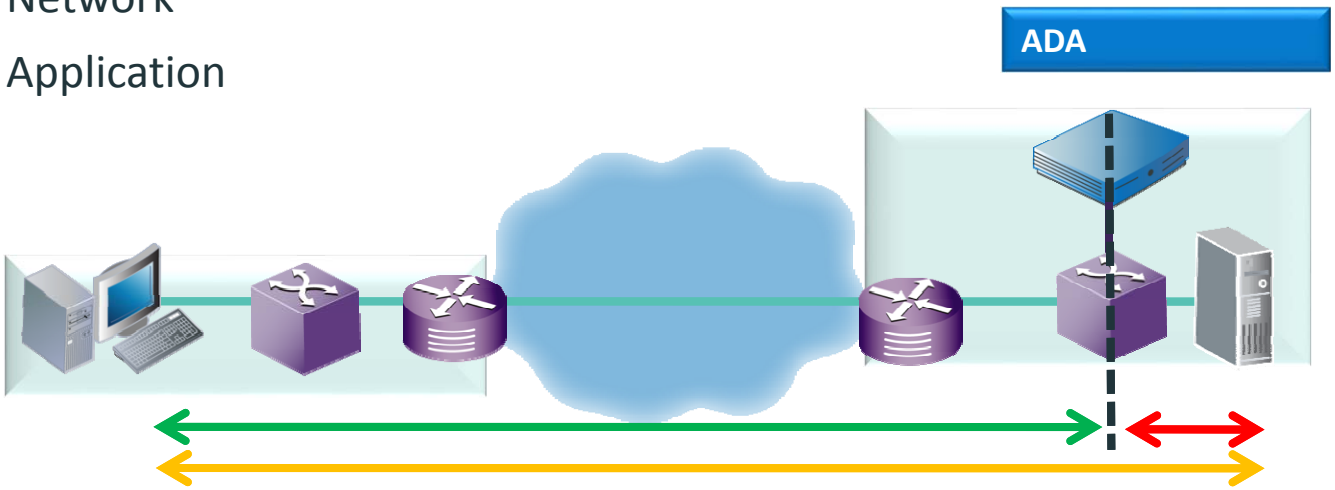
- ADA receives a concurrent copy of the frame
- Note: Alternate method of using a TAP

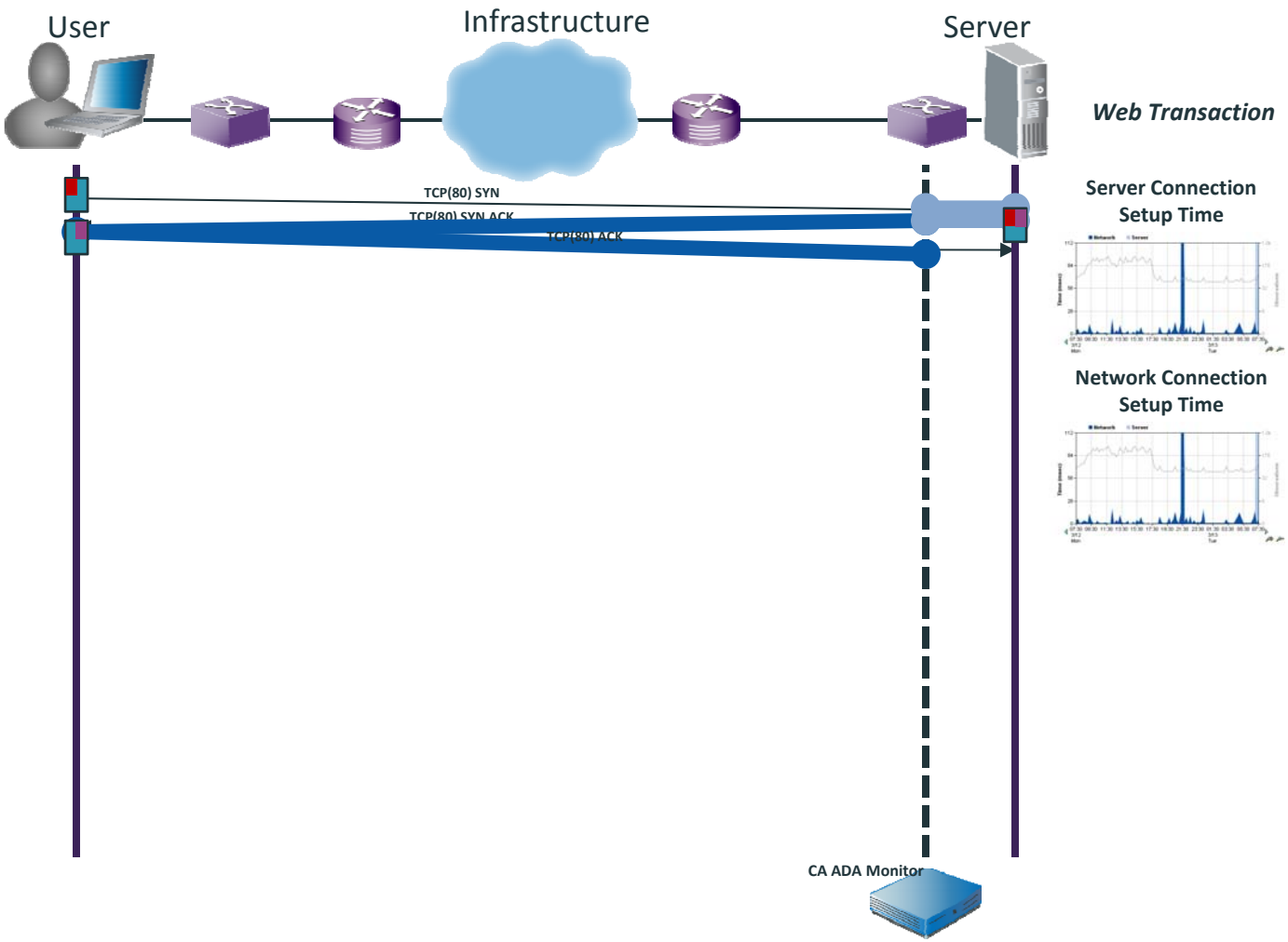


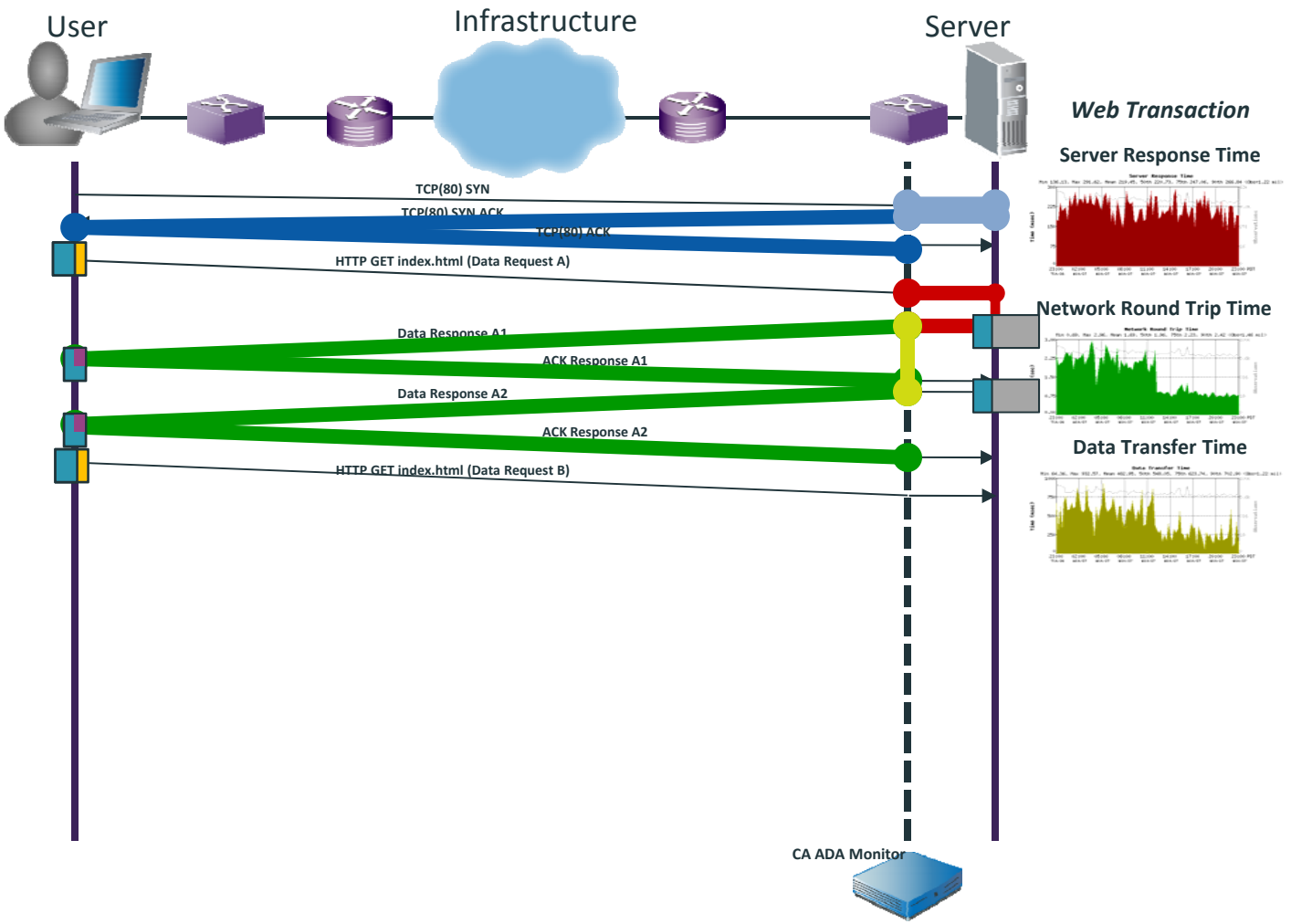
Location, Location, Location

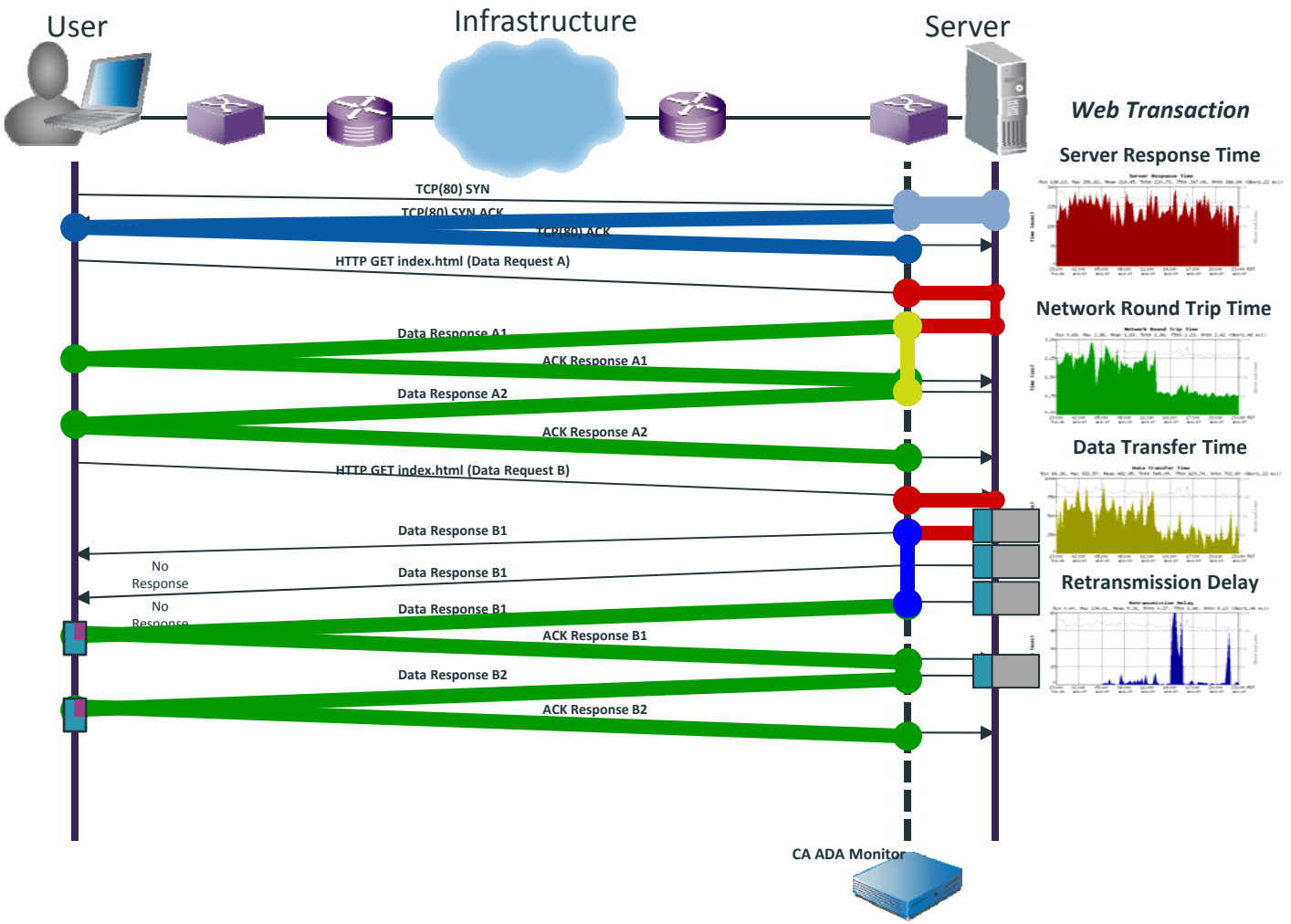
ADA uses its position next to the server to break response times into the basic components:

- Server
- Network
- Application









Response Time Insight

Server Response

+

Data Xfer

+

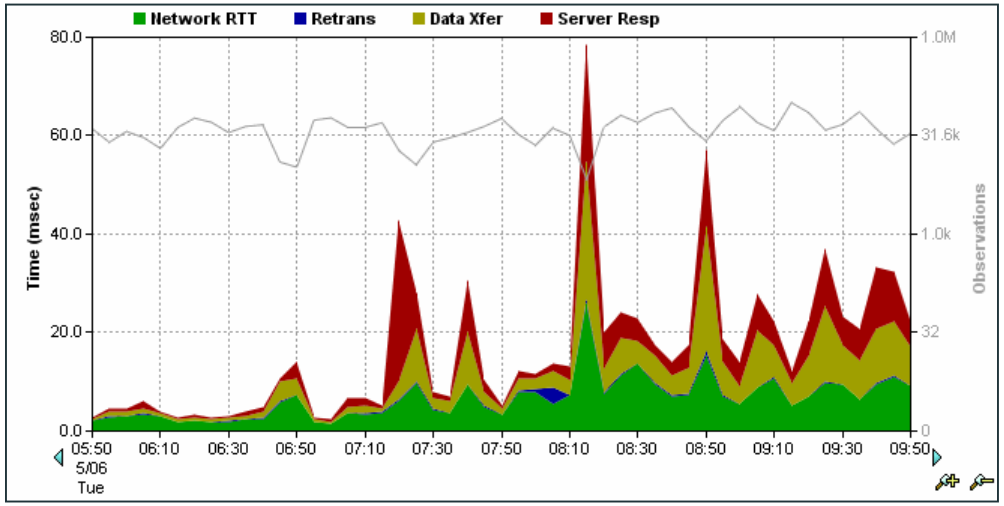
Retrans. Delay

+

Network RTT

=

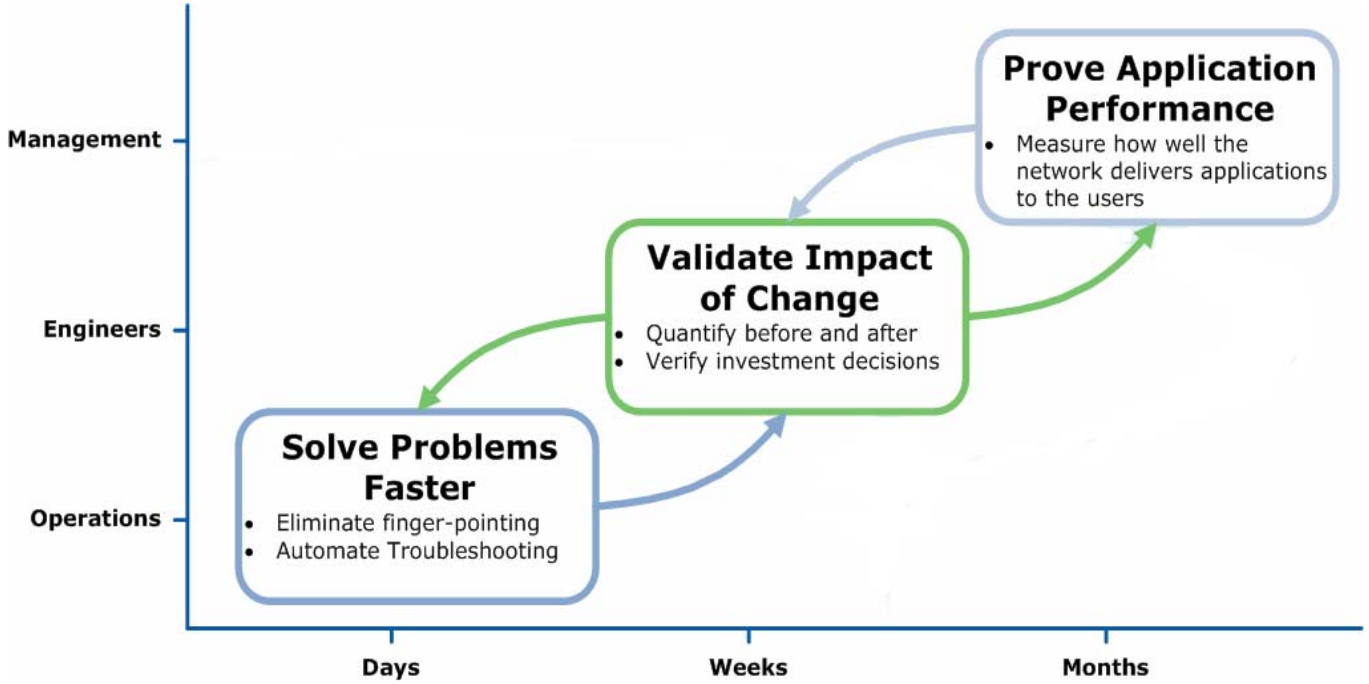
Total Transaction Time



Gray line represents observed transactions over given time interval

Average time for a TCP request to be fulfilled

Value of End-to-End Monitoring



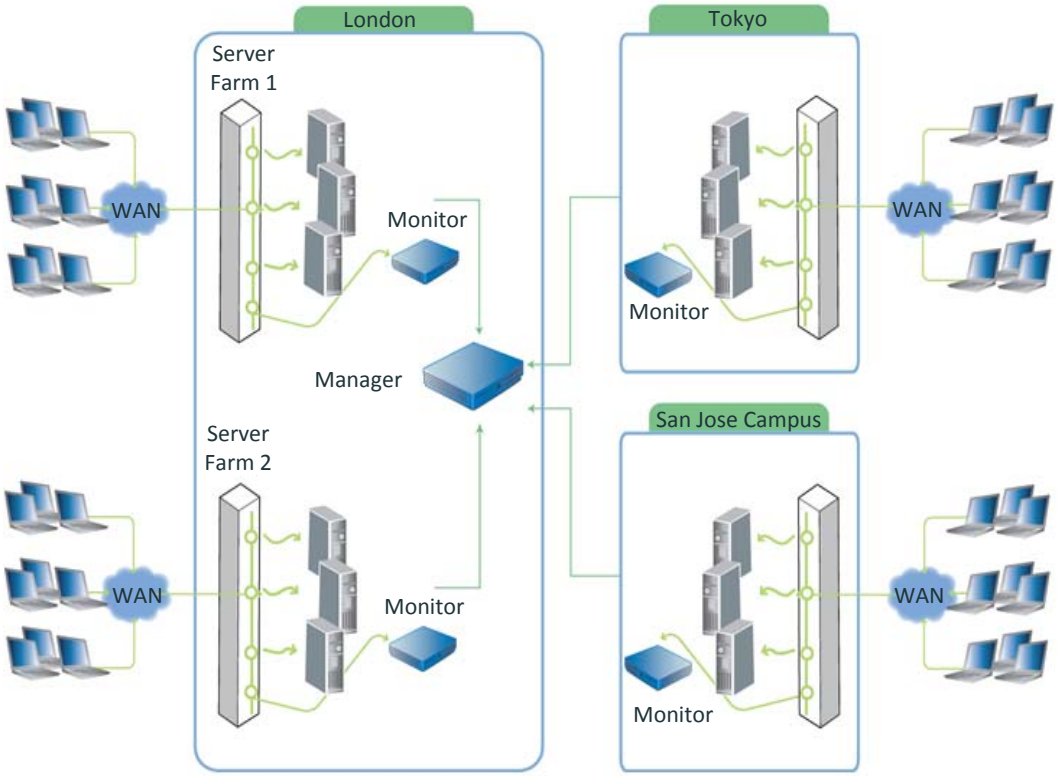
ca technologies

Architecture

Architecture & Data Feeds

- ADA supports a wide variety of data collection.
  - Monitors
    - Physical or Virtual
    - Single or Multi Port (Multi-port provides CEM TIM collector capability)
  - ADA system that accepts metrics from 3rd party devices
    - GigaStor
    - Cisco WAE & Riverbed Steelhead (Wan Optimization)
    - Cisco NAM (Switch Module)

Collection Options



Manager = Console
Monitor = Collector

ca technologies

Placement – where and why

Placement Best Practices

- Data Centre centric
- Connect to the Access Layer
- Monitor as many layers of the application as possible
- Monitor pre and post LB (VIP and RIP)
- VM Monitor only for traffic between guest servers in the same host
- Use MTP and filters to reduce traffic

Configuration

Configuration Best Practices

- Use MTP and filters to reduce traffic
- Have a clean SPAN
- Make sure you monitor what is important

Data Analysis

Eight Metrics that Matter Most

- Server
  - ✓ Unresponsive Sessions
  - ✓ Refused Sessions
- Application
  - ✓ Server Response Time
  - ✓ Application Turns
- Network
  - ✓ Packet Loss
  - ✓ Latency
- Additional Metrics
  - ✓ Server Connection Time
  - ✓ Data Transfer Time

    Network = Driven high due to increases in packet loss and/or latency
    Server = Driven high due to reduced TCP Receive Window causing delays
    Application = Pause after beginning of DTT while app retrieves data
    Averaged Volumes =

Key Server Metrics

TCP/IP Connection Setup

- The application cannot transfer data until a successful TCP connection has been established.
- If the session setup does not complete, the application transaction (or measurement of the transaction) never begins.
- Unfulfilled TCP/IP Sessions
  - ✓ Unresponsive Sessions
  - ✓ Refused Sessions

Session Setup

Unfulfulled TCP Sessions

- Unresponsive Sessions
  - ✓ An *unresponsive session* occurs when a connection request was sent, but the server never responded.
- Refused Sessions
  - ✓ A *refused session* occurs when a connection request was explicitly rejected by the server during the three-way handshake.

Key Application Metrics

Delivery

- Server Response Time (SRT)
  - ✓ Can be impacted by server performance as well as application architecture dependencies.
- Server Response Time Observations
  - ✓ Number of Application Turns
  - ✓ 1 Command + 1 Response = 1 Turn

Delivery

- Server Response Time (SRT)
  - ✓ This shows the amount of time a server takes to start responding to a request made by a client.
- Application Turns
  - ✓ The number of application turns can be determined by using the Observation (Obs) count inside the Server Response Time view.
    Observations = The number of times a particular metric is observed
    One SRT Observation = One Command/Response Combination

Server Response Time

Application Performance

- When SRT increases: check the performance metrics for all application dependencies
  - ✓ The UIM/ADA integration shows this server dependency view by default.
  - ✓ Copy the Server showing increased SRT and paste it into the “network” filter in ADA

    Clear all other filters but keep the same timeframe

    Note any increase in any of the key metrics identified in this presentation

    This would be a clear indication that an application dependency is causing the increased SRT

Server Response Time

Server Performance

- It is possible that the server showing the increased SRT is the reason for the performance degredation
- Any correlation between Server Connection Time (SCT) and SRT is a clear indication that the degredation is the result of the server's internal (or its virtual host's) performance
- Leverage UIM to dive deeper into why the server may be experiencing performance issues

Network Issues

There are only two things on a network that impact end-user performance

- Packet Loss
- Latency

Packet Loss

Two ways to display End-to-End Packet Loss in ADA

- Retransmission Delay
  - Pro – Minimally Impacted by Packet Duplication
  - Con – Harder to Quantify / Threshold
- Packet Loss Percentage
  - Pro – easier to quantify good vs. bad thresholds
    - Critical > 0.5%
    - Major > 0.25%
    - Minor > 0.05%
  - Con – Duplicate Packets Create High Impact for this Metric
    - Production Servers > 20% Packet Loss
    - Production Networks > 20% Packet Loss

Types of Packet Loss

There are two classifications for types of Packet Loss

- Errors
  - ✓ Data Corruption
- Discards
  - ✓ Capacity Issues

While packet loss can be measured end-to-end, errors and discards must be measured hop-by-hop.

- SNMP

Errors

Errors are the result of corrupted data

- Hardware
  - ✓ Transmitting NIC/Port
  - ✓ Receiving NIC/Port
  - ✓ Duplex Mismatch
- Cabling
  - ✓ Length
  - ✓ Condition
    - Crimp
    - Corrosion
  - ✓ Electromagnetic Interference
    - Noise

Discards - Capacity Issues

Inbound Discards

- System Unable to Process Packets
  - ✓ CPU
  - ✓ Memory
  - ✓ I/O
- Often Related to Packets per Second
  - ✓ Data Rate of about 10mb/s
    - 800 packets x 1518 bytes / second
    - 19,000 packets x 64 bytes / second

Outbound Discards

- System Unable to Offload Packets
- Lack of Bandwidth or Priority
  - ✓ Overloaded Interface Queue

Sources of Network Latency

There are five sources of delays associated with NRTT:

- Serialization Delay
  - ✓ Generally most significant on interface speeds below 10mbs
  - ✓ Minimal delays associated with minimum packet sizes
- Queuing Delay
  - ✓ Offers potential significant delay only when congestion exists
- Distance Delay
  - ✓ Distances can be estimated using Internet travel map applications
- Routing/Switching Delay
  - ✓ AKA: Forwarding Delay
- Protocol Delay

Measuring Latency

NRTT

Serialization + Queuing + Distance + Forwarding + Protocol = User Experience

NCT

- Queuing + Distance = Infrastructure Delivery
  - ✓ **Forwarding** < 3ms round trip
  - ✓ Minimal **Serialization** (0.3ms per T-1 hop)
  - ✓ No TCP Delay ACK **Protocol Delay**
  - ✓ Measure with LAN segments to avoid Wireless Protocol Delays

Priority of Network Metrics

Primary Metrics

- Packet Loss (end-to-end)
  - ✓ Errors
  - ✓ Inbound Discards
  - ✓ Outbound Discards
- Latency (end-to-end)
- Jitter (if video and/or VoIP is involved)

Priority of Network Metrics

Secondary Metrics

- Packets per Second
- Utilization
  - ✓ CPU
  - ✓ Memory
  - ✓ I/O (Read/Write)
  - ✓ Link (bandwidth)
- Latency (Device)
  - ✓ When routers and switches start to get busy, they may respond slower to Pings sent to them, than they do for traffic passing through them

Other Important Metrics

Server Connection Time (SCT)

Impacted only by the monitored server

Data Transfer Time (DTT)

Impacted by various conditions

Server Connection Time

Server Connection Time (SCT)

The time from the initial SYN packet being received from the client until the server sends out the first SYN/ACK.

Unlike SRT, Server Connection Time is not dependent upon backend services or any application architecture.

- High SCT indicates a server kernel level response time issue with that server.
- Use ADA to see if a substantial increase in the number of sessions hosted by the server has occurred.
- Use UIM/PM to investigate internal processes, memory and I/O functions of that server to see why the server is slow to respond.

Note that in an virtual instance, high SCT can be caused by either a delay on the server instance or a delay by the virtual host.

Data Transfer Time

The Data Transfer Time (DTT) metric can be impacted by any of these components.

- Network
- Application
- Server (acting as a Client)

Network Impact to DTT

Increases in packet loss or latency often impact Data Transfer Time (DTT)

- Whenever Data Transfer Time aligns with spikes in Retransmission Delay and/or Network Round Trip Time, it is safe to infer that the network is the cause of the increased DTT.

Application Impact to DTT

Data Transfer Time

- The majority of the time increases in DTT are the result of:
  - Increase in the size of the data delivered (1KB, 1MB or 100MB)
  - DTT uses an average of response sizes along with an average of the time to complete those responses
- Rarely does DTT increase as the result of application performance
  - However, occasionally an application will begin a data transfer and pause before delivering the remainder of the content
    - Web traffic delivering header information prior to completing the collection of the dynamic content from its backend processes

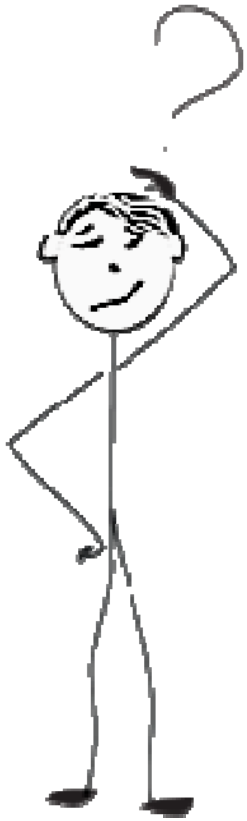
Server (Receiver) Impact to DTT

Data Transfer Time

DTT can be impacted when the receiver's TCP Window drops low enough to cause delays in the transfer process. Typically, this may occur if the TCP Receive Window drops below the MSS value of the given session.
When this condition occurs, the server in question is acting as a client. It is receiving data from its dependencies and is having difficulty processing that data at the same rate as it is being delivered.

Q & A

Questions?



ca technologies

Todor Kardjiev
Principal Consultant Technical Sales - EMEA
Todor.Kardjiev@ca.com