1 Metrics

Metrics include the performance and utilization data captured by the appliance directly from network traffic in real time, as well as information derived from traceroutes, BGP, DNS, and a WHOIS database in the appliance. Metrics provide insight into the behavior of individual IP addresses or the aggregate behavior of groups.

The following topics are discussed.

Metric Concepts

The following topics describe important metric concepts and metric-related terms that appear in tables, charts, and insights:

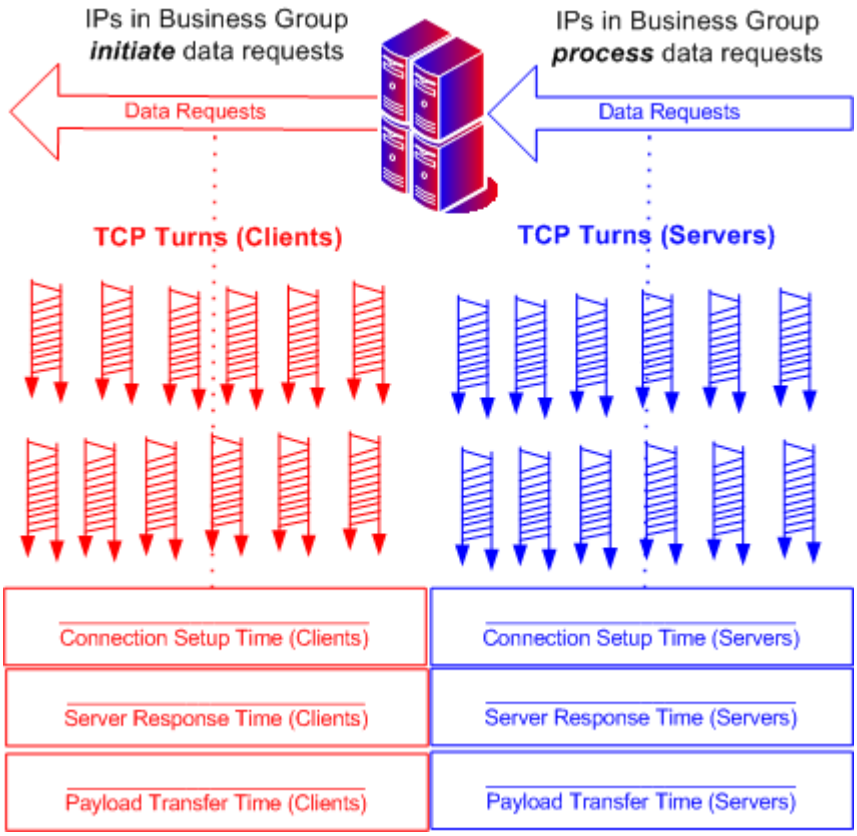
- TCP Client
- Inbound / Outbound
- Response Time
  - Server Response Time
  - User Response Time
  - Server and User Response Times, Clients and Servers
- Server
- TCP Client
- TCP Connection
- TCP Server
- TCP Turn

Client

An IP that sends the data request in a TCP Turn. When you see a metric that reads "[*metric*] (Clients)," the metric refers to all IPs in the Group that sent TCP data requests within the current Project Time window.

A Group can include both Clients (IPs that send data requests) and Servers (IPs that receive and process data requests), as illustrated in the following figure.

Figure 1-1 Metrics for "(Clients)" and "(Servers)" in a Group



Inbound / Outbound

Traffic direction relative the appliance and/or to a group. These terms have different meanings in Utilization Metrics and Network Performance Metrics.

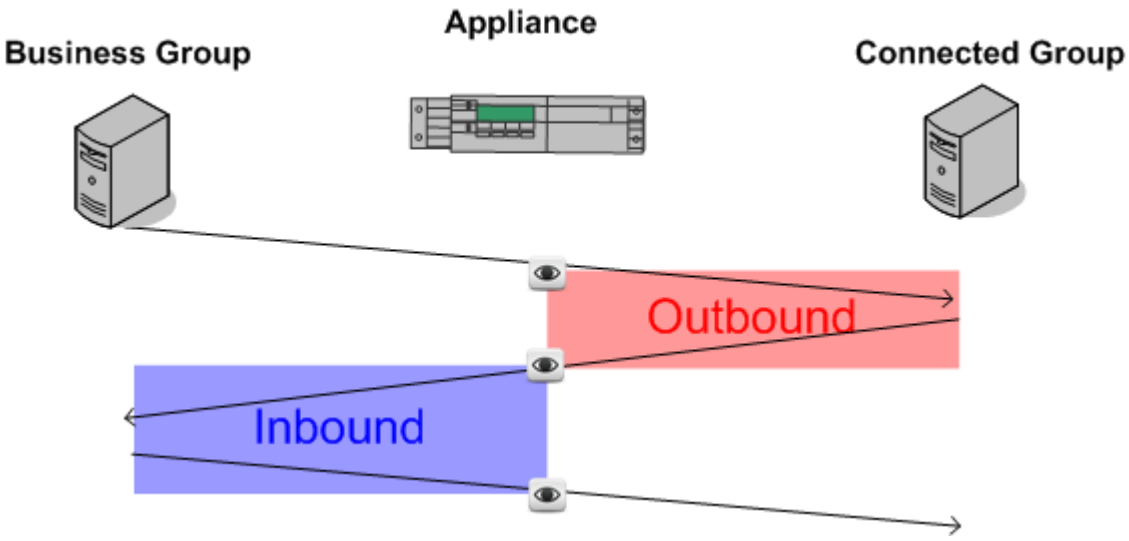
Utilization Metrics measure Inbound traffic (to the group) and Outbound traffic (from the group).This rule applies to all groups with the following exceptions.

- Total Traffic
- ISP AS
- Peer AS

For more information about these exceptions, see Utilization Metric Sources on page usr-3-90.

For Network Performance Metrics, Inbound and Outbound are relative to the position of the appliance between groups. In the following figure, Inbound is `appliance<-->business_group`, while Outbound is `appliance<-->connected_group`.

Figure 1-2 TCP Network Metrics - Inbound/Outbound



Response Time

Key Concept—Application Performance Metrics collected by the appliance are based on TCP Turns. This means that Response Time metrics are TCP-based and are different from the “response times” as observed at the application layer or as experienced by an end user.

The following topics discuss Response Times in detail:

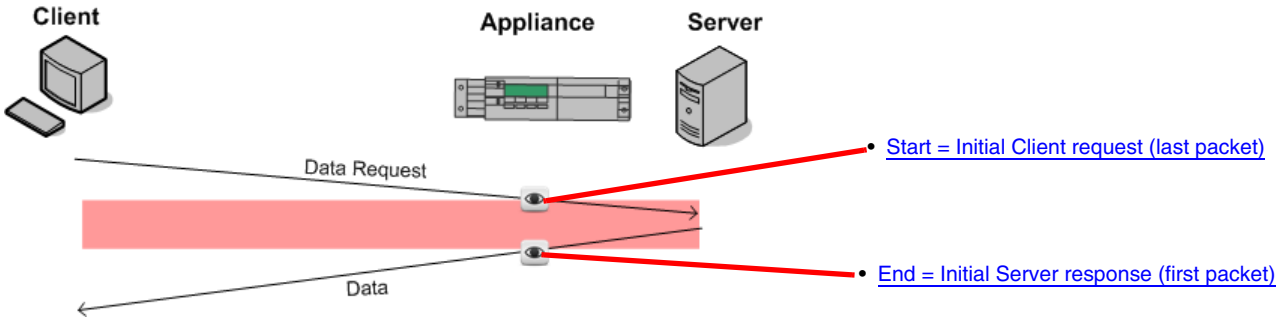
- Server Response Time
- User Response Time
- Server and User Response Times, Clients and Servers

Server Response Time

The Server Response Time measures how quickly a Server can respond to a data request (*ip_gets_request* to *ip_sends_response*). The Server Response Time for an individual turn is measured as follows:

- Start = Initial Client request (last packet)
- End = Initial Server response (first packet)

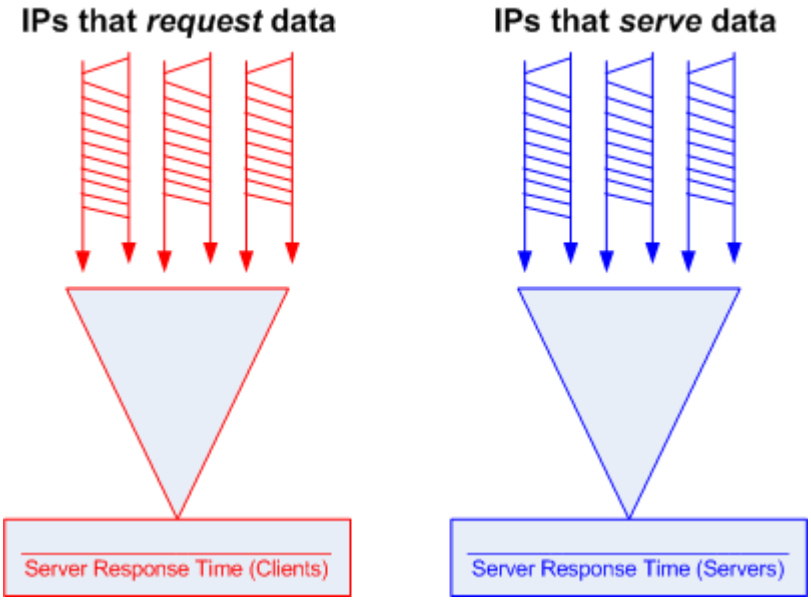
Figure 1-3 Server Response Time for an Individual Turn



For each Group (Business Group, Application, IP, etc.) for which metrics are collected, the appliance calculates two averages:

- Server Response Time (Clients) = All IPs in the Group that requested data
- Server Response Time (Servers) = All IPs in the Group that served data

Figure 1-4 Server Response Times for "(Clients)" and "(Servers)" in a Group



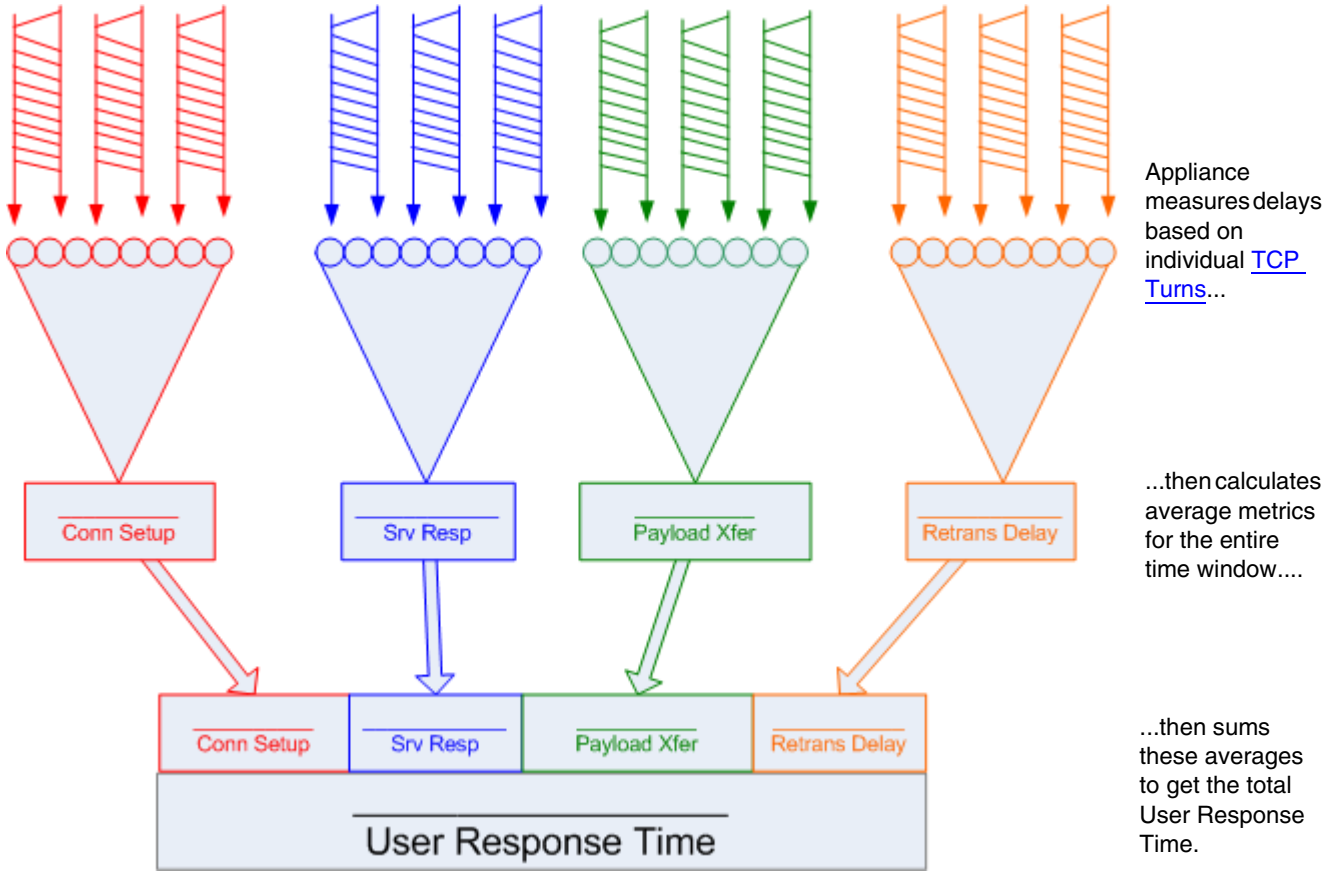
The appliance calculates a separate averages based on two sets of TCP Turns:

- Turns in which member IPs are Clients
- Turns in which member IPs are Servers

User Response Time

User Response Time is an estimate of how quickly a Client will receive the payload for an individual data request (*ip_requests_data* to *ip_gets_full_requested_payload*). This metric is a sum of the averages of the following four metrics: Connection Setup Time [msec] + Server Response Time [msec] + Payload Transfer Time [sec] + Retransmission Delay [msec].

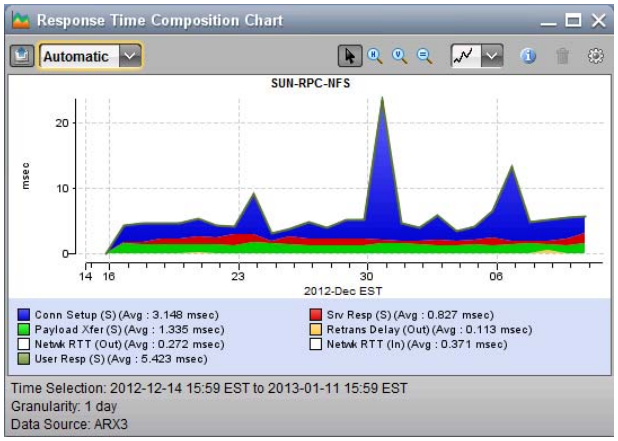
Figure 1-5 User Response Time: Sum of Averages for Four Metrics



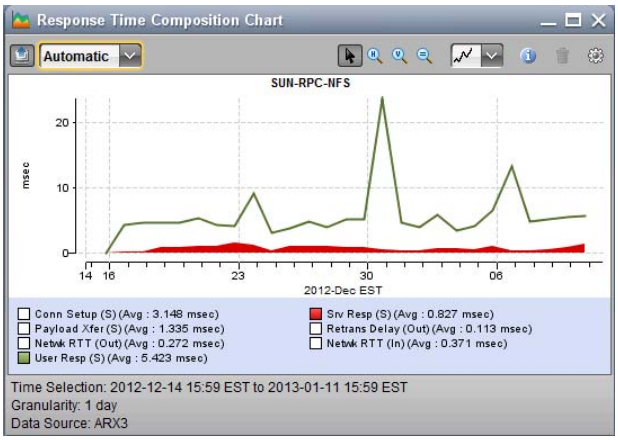
User Response Time estimates the average time for an application to process a TCP data request, and thus is a good measure of end-user response time.

A good way to view the relationship of Server Response Time and User Response Time is in a Response Time Composition Chart. If you remove the Round Trip Times from the chart, you have the User Response Time. If you remove Payload Transfer, Connection Setup, and Retransmission Delay, you can see the relation between the User and Server Response Times.

Figure 1-6 Server and User Response Times in the Response Time Composition Chart



User Response Time with all four component metrics



User Response Time and Server Response Time only

Server and User Response Times, Clients and Servers

You can think of User Response Time and Server Response Time as the average duration of all data requests/responses as seen from two perspectives:

- Server Response Time = How quickly a Server responded to the data requests it received (on average), from *ip_gets_request* to *ip_sends_response*
- User Response Time = How quickly a Client received the payloads for the data requests it sent (on average), from *ip_requests_data* to *ip_gets_full_requested_payload*

While User and Server Response Times refer to the perspective—*where* the "response time" is experienced and estimated—Clients and Servers refer to *which* IPs are included in the Response Time average:

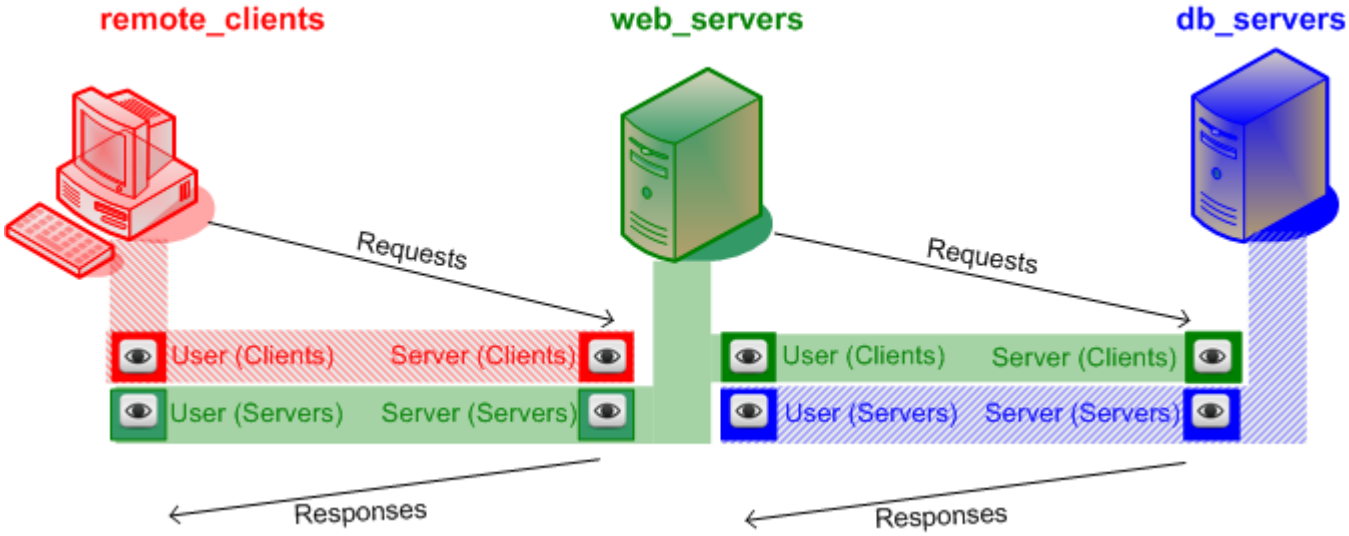
- Response Time for Clients (IPs that request data)
- Response Time for Servers (IPs that serve data)

Thus, we have:

- User Response Time (Clients) = Response times from the *user request* perspective; average for all IPs that *request* data
- User Response Time (Servers) = Response times from the *user request* perspective; average for all IPs that *serve* data
- Server Response Time (Clients) = Response times from the *server response* perspective; average for all IPs that *request* data
- Server Response Time (Servers) = Response times from the *server response* perspective; average for all IPs that *serve* data

The following example shows three Business Groups: remote_clients request data from web_servers, and web_servers request data from db_servers. In this case, the web_servers act as both Clients and Servers.

Figure 1-7 User and Server Response Times, Clients and Servers: Example

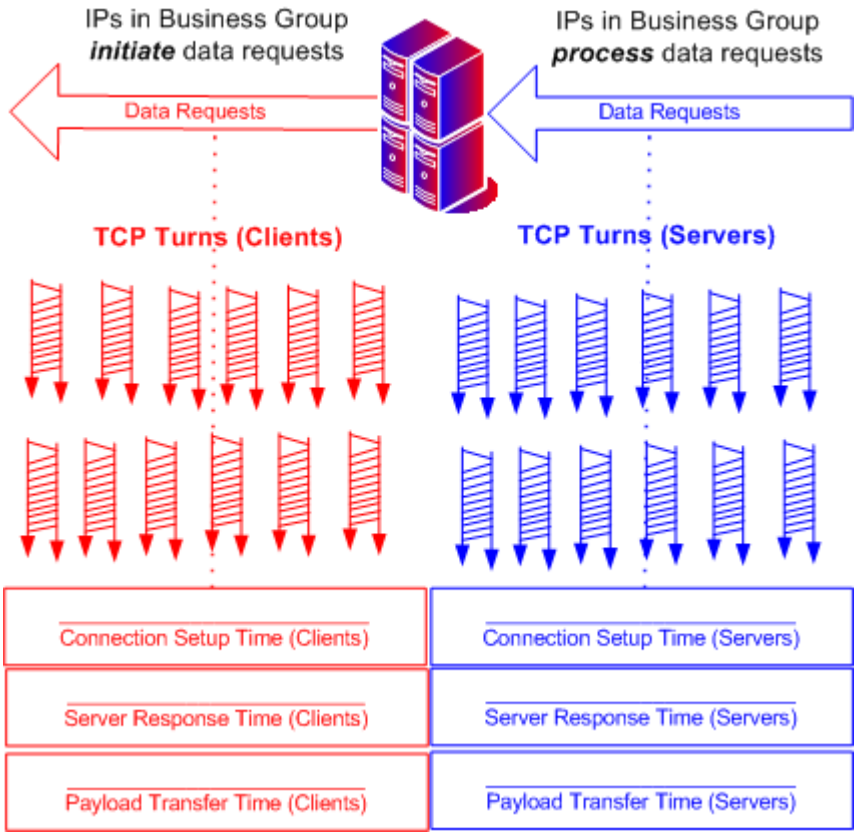


Server

An IP that receives and processes the data request in a TCP Turn. When you see a metric that reads "[*metric*] (Servers)," the metric refers to all IPs in the Group that received and processed TCP data requests within the current Project Time window.

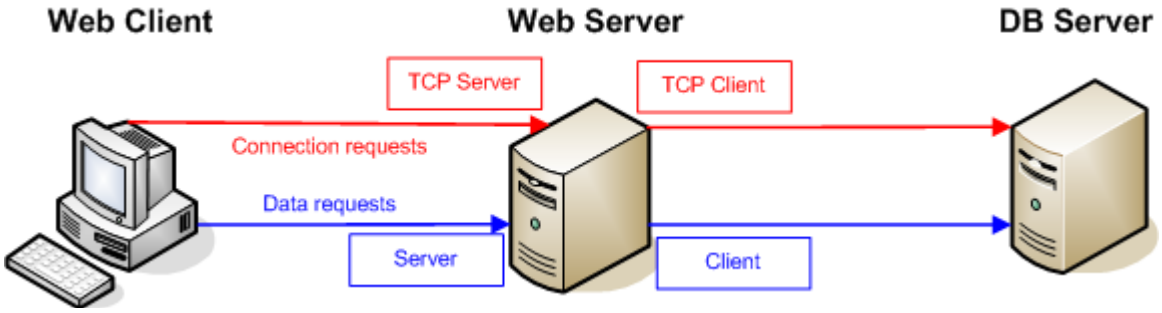
A Group can include both Clients (IPs that send data requests) and Servers (IPs that receive and process data requests), as illustrated in the following figure.

Figure 1-8 Metrics for "(Clients)" and "(Servers)" in a Group



In some cases, an IP can act as a Client (requests data), Server (serves data), TCP Client (requests connections), and TCP Server (serves connections).

Figure 1-9 Business Group as TCP Server, TCP Client, Server, Client

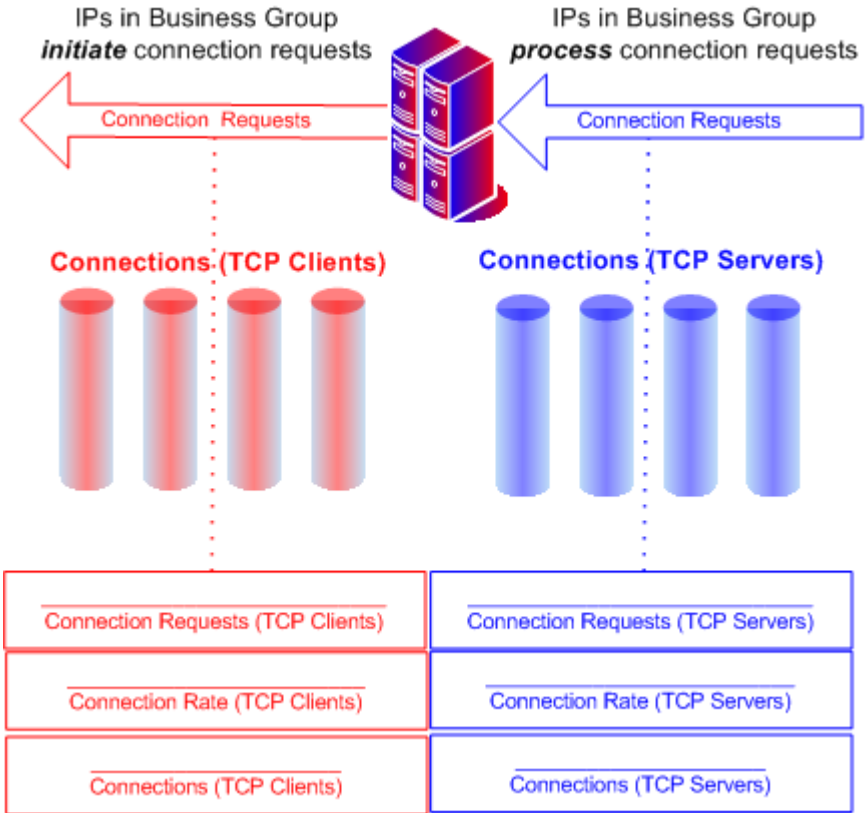


TCP Client

An IP that initiates a TCP Connection. When you see a metric that reads "[*metric*] (TCP Clients)," the metric refers to all IPs in the Group that sent TCP connection requests within the current Project Time window.

A Group can include both TCP Clients (IPs that send connection requests) and TCP Servers (IPs that receive and process connection requests), as illustrated in the following figure.

Figure 1-10 Metrics for "(TCP Clients)" and "(TCP Servers)" in a Group



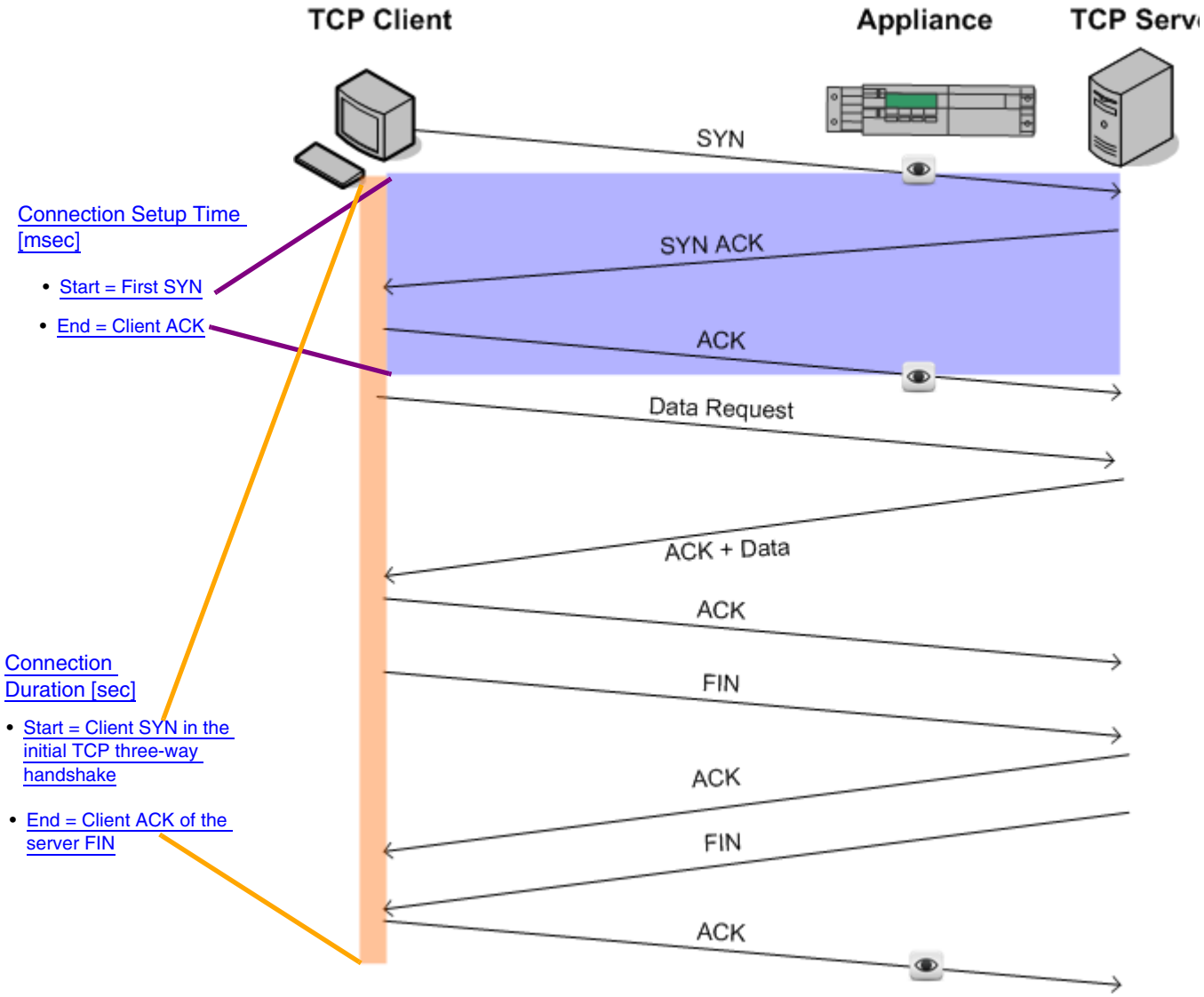
TCP Connection

A TCP connection observed by the appliance. All Application Performance metrics labelled "(TCP Clients)" and "(TCP Servers)" are based on TCP Connections and the role of IPs in these connections:

- TCP Clients refers to all IPs that requested connections during the current Project Time. For example, Connection Requests [#] (TCP Clients) is the total number of connection requests generated by IPs in the Group during the current Project Time.
- TCP Servers refers to all IPs that responded to connection requests during the current Project Time. Another way to define a TCP Server is: an IP that listens on a well-known port for connection requests.

TCP Connections usually include multiple TCP Turns; metrics labelled "(Clients)" and "(Servers)" are based on the turns for a specific Group. The following figure shows how the appliance measures two connection metrics using SYN and ACK packets.

Figure 1-11 Metrics for an Individual Connection

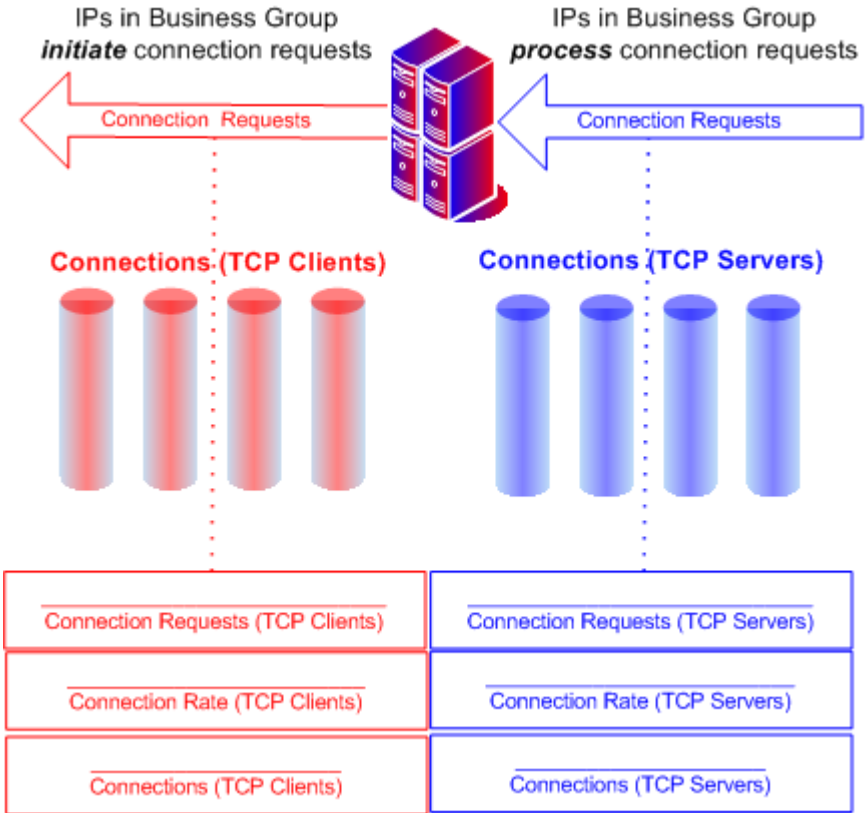


TCP Server

An IP that listens on a well-known port for TCP Connection requests. When you see a metric that reads "[*metric*] (TCP Servers)," the metric refers to all IPs in the Group that responded to TCP connection requests within the current Project Time window.

A Group can include both TCP Clients (IPs that send connection requests) and TCP Servers (IPs that receive and process connection requests), as illustrated in the following figure.

Figure 1-12 Metrics for "(TCP Clients)" and "(TCP Servers)" in a Group



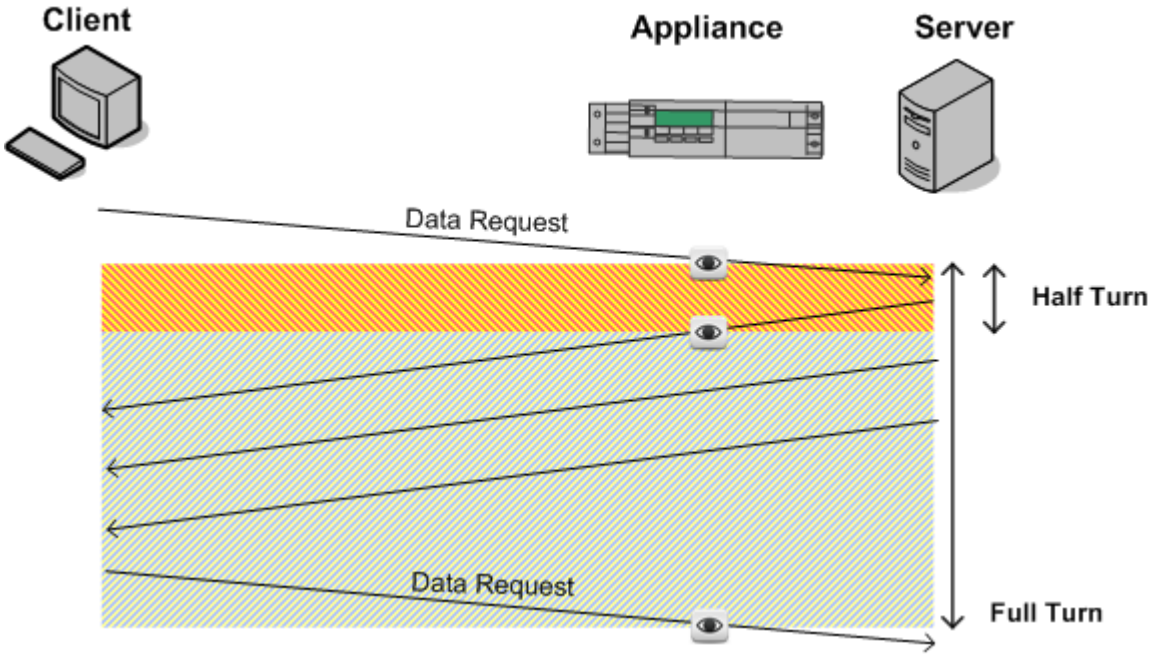
TCP Turn

One full TCP data request-response cycle. All Application Performance metrics labelled "(Clients)" and "(Servers)" are based on TCP Turns.

Many applications establish a persistent TCP connection that includes many individual request-responses. Turn metrics are based on a generic Layer-4 model and usefully approximate end-user experience and other aspects of application performance.

ARX calculates individual turns based on changes in the direction of data packets. Request N (client => server) is the turn start; response N (server => client) is the half-turn; request N + 1 (client => server) is the full turn.

Figure 1-13 TCP Turn Start and End, Based on Changes in Direction of Data Packets



ARX identifies data-request and data-response packets based on the following criteria:

- Data request = An packet from the Client to the Server that includes data (i.e., that has a non-zero payload)
- Data response = A packet with a non-zero payload, sent in response to a data request

When the appliance determines when specific turns start and end, it considers *data packets only*. It does not consider dataless packets such as zero-payload ACKs, SYNs, and FINs used for connection management and flow control. The appliance does consider dataless packets when calculating certain turn metrics, however.

Metric Definitions

The following topics define the metrics that the appliance collects and shows in tables, charts, and insights:

- Informational Metrics Definitions
- Application Performance Metric Definitions
- Network Performance Metric Definitions
- Utilization Metric Definitions
- Page Performance Metric Definitions

—

Informational Metrics Definitions

The following table describes the Informational metrics available in AppResponse Xpert. **Note—**Information metrics can be added and removed from tables, but cannot be displayed in charts.

Table 1-1 Informational Metrics

Metric	Definition
AS Description	The Autonomous System description. This information is stored in the WHOIS Database, which you can edit using the AS Information Manager. If the group does not appear in the WHOIS database, the field is left blank or displayed as Unknown.
AS Name	The Autonomous System to which the group belongs. Available for AS groups (ISP AS, Peer AS or Dest AS) and IP addresses only. This information is stored in the WHOIS Database, which you can edit using the AS Information Manager. If the group does not appear in the WHOIS database, the field is left blank or displayed as Unknown.
AS Number #	Autonomous System number of the group. Available for AS groups (ISP AS, Peer AS or Dest AS) and IP addresses. This information is stored in the WHOIS Database, which you can edit using the AS Information Manager. If the group does not appear in the WHOIS database, the field is left blank or displayed as Unknown.
Group	Group or drilldown description.
Group Definition	Group definition—for example, the set of member IPs for a Business Group or the assigned IP protocol and port(s) for a Defined Application.
Group Membership	The set of all Business Groups or Business Group Containers to which an IP address belongs.
Information	Additional information for a specific group, such as the DNS hostname for IP addresses and the Autonomous System name for AS groups (ISP AS, Peer AS, Dest AS).

Related Topics

- *WHOIS Database*
- *AS Information Manager*

Application Performance Metric Definitions

The following table describes the Application Performance metrics available in AppResponse Xpert.

Table 1-1 Application Performance Metrics

Metric Name	Definition
Client Reset Rate [*#/sec*] TCP Clients / TCP Servers	Average number of TCP Connections that terminated with a TCP client reset within the selected Project Time.
Connection Duration [*sec*] TCP Clients / TCP Servers	The average duration of all TCP Connections that ended within the current Project Time. The Connection Duration is measured as follows. *Under regular network conditions:* • Start = Client SYN in the initial TCP three-way handshake • End = Client ACK of the server FIN *If client traffic is not observed:* • Start = Server SYN-ACK in the initial three-way handshake • End = Server FIN *If the connection is reset without the standard TCP termination sequence:* • Start = Client SYN in the initial three-way handshake • End = Time when TCP reset is observed *If the connection times out:* • Start = Client SYN in the initial three-way handshake • End = 30 seconds later, if no SYN or FIN is observed **Note—**This metric is reported differently in charts and tables. For more information, see Special Handling of Long Connections and Turns on page usr-3-95.
Connection Rate [#/sec] TCP Clients / TCP Servers	Number of successful TCP Connections per second. The three-way handshake must be completed for the connection to be counted as successful.
Connection Request Rate [#/sec] TCP Clients / TCP Servers	Number of attempted TCP Connections per second. A connection request occurs when the client sends a SYN request to the server, regardless of whether or not the server responds.
Connection Requests [#] TCP Clients / TCP Servers	The number of attempted TCP Connections within the selected Project Time. An attempted connection occurs when the client sends a TCP SYN request to the server, regardless of whether or not the server responds.
Connection Setup Time [*msec*] TCP Clients / TCP Servers	The average time required to establish the initial three-way handshake for a TCP connection: • Start = First SYN • End = Client ACK

—

Table 1-1 Application Performance Metrics (Continued)

Metric Name	Definition
Connections [*#*] TCP Clients / TCP Servers	Total number of successful TCP Connections. The three-way handshake must be completed for the connection to be counted as successful.
Connections Failed [*#*] TCP Clients / TCP Servers	The number of connection requests to which servers in the group do not respond within the selected Project Time.
Connections Failed Rate [*#/sec*] TCP Clients / TCP Servers	Average number of failed TCP Connections per second.
Data Transfer Time [*sec*] Clients / Servers	The average time required to transfer data, for all TCP Turns that ended within the selected Project Time. The Data Transfer Time for an individual turn is measured as follows: • Start = Initial Server response (first byte) • End = Client ACK for last response packet **Note—**This metric is reported differently in charts and tables. For more information, see Special Handling of Long Connections and Turns on page usr-3-95.
Initial Application Response Time [*msec*] TCP Clients / TCP Servers	The average time it takes for a TCP Server to respond to an application-layer request. • Start = TCP Server receives application-layer request (such as HTTP GET) • End = TCP Server sends first byte in response to request This metric measures application-layer response times at the TCP Server. To calculate response times at the server host, regardless of network effects, AppResponse Xpert subtracts server-side round trip times observed in the network. The round-trip time is measured during the three-way handshake at the beginning of the TCP connection; it is the delay between the initial TCP SYN from client to server (observed at the appliance) and the responding TCP SYN/ACK from server to client (observed at the appliance). **Note—**AppResponse Xpert does not calculate this metric if the client does not send a connection request or if the appliance does not observe the request.
Payload [*KB*] Clients / Servers	The average size of data transfers from Servers to Clients, for all TCP Turns that ended in the selected time interval. Payload does not include retransmissions or zero-payload ACKs. **Note—**This metric is reported differently in charts and tables. For more information, see Special Handling of Long Connections and Turns on page usr-3-95.
Payload Transfer Time [*sec*] Clients / Servers	The average time to deliver a good payload per TCP Turn. This metric is a subset of the existing Data Transfer Time where Payload Transfer Time = Data Transfer Time [sec] - Retransmission Delay [msec]

Table 1-1 Application Performance Metrics (Continued)

Metric Name	Definition
Server Reset Rate [*#/sec*] TCP Clients / TCP Servers	The average number of TCP Connections that were terminated with a TCP reset by a TCP Server.
Server Response Time [*msec*] Clients / Servers	The average time for a Server to respond to a data request. The Server Response Time for an individual turn is measured as follows: • Start = Initial Client request (last packet) • End = Initial Server response (first packet) For more information about how this metric is calculated, see Response Time.
Time To First Byte [*msec*] TCP Clients / TCP Servers	The average time for a TCP Server to send the first data byte in a TCP Connection. The Time To First Byte for an individual connection is measured as follows: • Start = Client SYN (first packet in initial three-way handshake) • End = Server data response (first packet)
Turn Rate [*#/sec*] Clients / Servers	Average number of responses from Turn Servers.
Turns (Clients) [*#*] Clients / Servers	Total number of TCP Turns.
User Response Time *[sec]* Clients / Servers	The average time it takes for an application to complete a TCP Turn. This metric is a sum of other metrics: User Response Time = Connection Setup Time [msec] + Server Response Time [msec] + Payload Transfer Time [sec] + Retransmission Delay [msec] This metric is a good measure of end-user response time when you look at a Response Time Composition Chart. If you remove the Round Trip Times from the chart, you have the User Response Time. For more information about how this metric is calculated, see Response Time.

Network Performance Metric Definitions

The following table describes the Network Performance metrics available in AppResponse Xpert.

Table 1-1 Network Performance Metrics

Metric	Definition
Packet Loss [%] Inbound / Outbound	Percentage of TCP packets that were retransmitted. A high percentage of retransmissions often indicates packet loss in the network, but retransmissions may also be caused by other end-to-end TCP effects. This metric is a measure of lost packets as a percentage of all packets observed for the Group of interest. Unlike packet loss reported on routers, which measures loss on one link only, this metric measures loss on all links where packets for the Group of interest were observed.
Packet Retransmission Rate [*#/sec*] Inbound / Outbound	Average number of TCP packets retransmitted per second.
Retransmission Delay [*msec*] Inbound / Outbound	Average delay per TCP Turn due to packet retransmissions. Retransmission Delay is calculated as Data Transfer Time [sec] multiplied by an empirically-derived function of packet loss. **Note—**This metric is reported differently in charts and tables. For more information, see Special Handling of Long Connections and Turns on page usr-3-95.
Retransmission Rate [*kbits/sec*] Inbound / Outbound	Average amount of retransmitted TCP data per second over the selected time interval.
Round Trip Time [*msec*] Inbound / Outbound	Average round trip time between the appliance and the TCP Client, over all TCP Connections. The Round Trip Time for an individual packet is measured as follows: • Start = TCP data packet from Server (observed by appliance) • End = Corresponding ACK from Client (observed by appliance) See also Note about Default Calculation of Round Trip Times on page 1-1-22.

Note about Default Calculation of Round Trip Times

Release 8.5 introduced a change in the default method for calculating the Round Trip Time [msec] metric. The default method considers only immediate ACKs and ignores delayed ACKs and piggybacked ACKs, resulting in round trip times that more closely reflect network propagation delay. The difference in round trip times (compared to previous releases) can be especially apparent for highly chatty applications. The trade-off of the new method is that you might see no Round Trip Time metrics at all for some highly chatty applications.

If you want to revert to the old method of calculating round trip times, do the following:

1) Using a SSH-enabled command line program such as putty, log in to the appliance as a user with admin privileges.

2) Enter the following command in the CLI:

```
setNgfestats COUNT_ONLY_IMMEDIATE_RTT=0
```

To switch back to the new method later, enter the following command in the CLI:

```
setNgfestats COUNT_ONLY_IMMEDIATE_RTT=1
```

Utilization Metric Definitions

The utilization metrics represent data captured from the IP headers of any IP traffic on the network.

Table 1-1 Utilization Metrics

Metric	Definition
95th Percentile Throughput *[kbits/sec]* Inbound / Outbound	95 percent of the time, throughput is lower than the specified rate. This metric is commonly used to evaluate peak network utilization. To determine the 95th Percentile Throughput, the appliance discards the top 5 percent of all throughput rates; the highest remaining rate is the 95th percentile. All calculations are based on 5-minute data granularity.
Goodput *[kbits/sec]* Inbound / Outbound	Effective transfer rate for all TCP payload data over all TCP connections. Goodput excludes retransmitted packets and TCP packets without data (such as SYNs or ACKs). **Note—**Although this metric is listed under Utilization Metrics, Goodput is based on TCP Connection data.
Packet Payload *[#/]* Clients / Servers	Average number of TCP data packets required to transfer data in all TCP Turns. Packet Payload excludes retransmissions and TCP packets without data (such as SYNs or ACKs). For information about how this metric is recorded and reported in charts and tables, see Special Handling of Long Connections and Turns on page usr-3-95.
Packet Size *[# bytes]* Inbound / Outbound Inbound and Outbound	Average number of bytes per packet (all layers)
Packet Throughput *[#/sec]* Inbound / Outbound Inbound and Outbound	Average number of packets transmitted per second.

Table 1-1 Utilization Metrics

Metric	Definition
Packet Traffic *[# packets]* Inbound / Outbound Inbound and Outbound	Total number of packets observed over all connections.
Payload [*# kB]* Clients / Servers	Average size of data transfers for all TCP Turns that ended within the selected time window. Note the following: • Payload excludes retransmissions and TCP packets without data, such as SYNs and ACKs. • Payload is always server-to-client, whether the metric is for Clients (servers –> clients_in_group) or Servers (servers_in_group –> clients). For information about how this metric is recorded and reported in charts and tables, see Special Handling of Long Connections and Turns on page usr-3-95.
Throughput *[MB/sec]* Inbound / Outbound Inbound and Outbound	Average data transmission rate, in raw Megabytes (headers + overhead + payload).
Traffic *[MB]* Inbound / Outbound	Total amount of data transmitted, in raw Megabytes (headers + overhead + payload).

Page Performance Metric Definitions

The following table describes the available web page metrics in AppResponse Xpert. You can view these metrics in the Web Applications Table and the "Individual Page Views" Insight.

Table 1-1 Web Page Metrics

Metric	Description
% HTTP 100s	Percentage of HTTP responses with status codes between 100-199
% HTTP 200s	Percentage of HTTP responses with status codes between 200-299
% HTTP 300s	Percentage of HTTP responses with status codes between 300-399
% HTTP 400s	Percentage of HTTP responses with status codes between 400-499
% HTTP 500s	Percentage of HTTP responses with status codes between 500-599
% Slow Pages (Clients)	Percentage of page views with page times that exceed a slow threshold for pages requested by Clients
% Slow Pages (Servers)	Percentage of page views with page times that exceed a slow threshold for pages requested from Servers
% Slow Pages	Percentage of page views with page times that exceed a slow threshold
HTTP 100s	Number of HTTP responses with status codes between 100-199
HTTP 200s	Number of HTTP responses with status codes between 200-299
HTTP 300s	Number of HTTP responses with status codes between 300-399
HTTP 400s	Number of HTTP responses with status codes between 400-499
HTTP 500s	Number of HTTP responses with status codes between 500-599
HTTP Responses	Number of HTTP server responses to HTTP client requests
Page Size [kB]	Average size of pages sent
Page Time (Clients)	Average time from first-to-last observed HTTP response in a page view for pages requested by Clients
Page Time (Servers)	Average time from first-to-last observed HTTP response in a page view for pages requested from Servers
Page Time	Average time from first-to-last observed HTTP response in a page view
Page Traffic [MB]	Sum of all HTTP response messages sent
Page View Rate	Average number of page views per minute
Page Views (Clients)	Number number of page views (page requested by Client IP)

Table 1-1 Web Page Metrics (Continued)

Metric	Description
Page Views (Servers)	Number of page views (page requested by Server IP)
Page Views	Number of requests to load a page
Request Size [kB]	Average size of HTTP request messages sent
Request Traffic [MB]	Sum of all HTTP request messages sent
Slow Page Rate (Clients)	Average rate of page views with page times that exceed a slow threshold for pages requested by Clients
Slow Page Rate (Servers)	Average rate of page views with page times that exceed a slow threshold for pages requested from Servers
Slow Page	Average rate of page views with page times that exceed a slow threshold
Slow Pages (Clients)	Number of page views with page times that exceed a slow threshold for pages requested by Clients
Slow Pages (Servers)	Number of page views with page times that exceed a slow threshold for pages requested from Servers
Slow Pages	Number of page views with page times that exceed a slow threshold