3 HTTP Server Performance Comparisons:

Throughput Comparison of 3 Servers

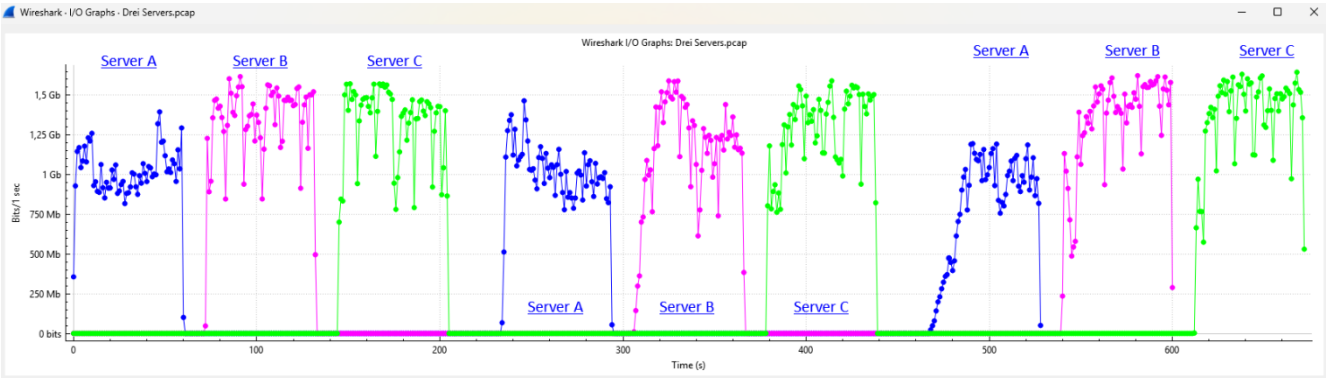
The figure below shows the measured downlink throughput of 3 different servers i.e. Server A (IPv6), Server B (IPv6) and Server C (IPv4) derived from Wireshark I/O graph analysis using a 1-second averaging interval.

The result shows that server B and sever C consistently achieves higher and more stable throughput with peak rates reaching approximately 1.5 Gbps during 60s test interval.

In contrast, server A exhibits slightly lower average throughput and greater variability. In particular, during the 3rd download interval, server A shows a noticeably slower throughput ramp-up phase. The steady-state throughput of server A remains closer to ~ 0.9 - 1.2 Gbps, compared to the higher plateau reached by server B and C.

All measurements were conducted sequentially, as indicated by the non-overlapping activity periods, ensuring that the observed differences are not caused by traffic concurrency.

The slower ramp-up and lower sustained throughput observed by server A are therefore likely attribute to server-side behavior transport layer configuration (e.g. TCP initial Congestion Window, Congestion Control Algorithm) rather than access network (5G/4G air interface) limitation.



Server Performance Comparison

Example: Bytes-in-Flight Comparison (3rd Download Session for Server A, B and C)

The below figure shows the tcp.analysis.bytes_in_flight for the third download session for Server A, B and C plotted with a 1-second interval.

Bytes-in-Flight reflects the amount of unacknowledged TCP data in the network and is a direct indicator of congestion window growth and utilization of the available bandwidth.

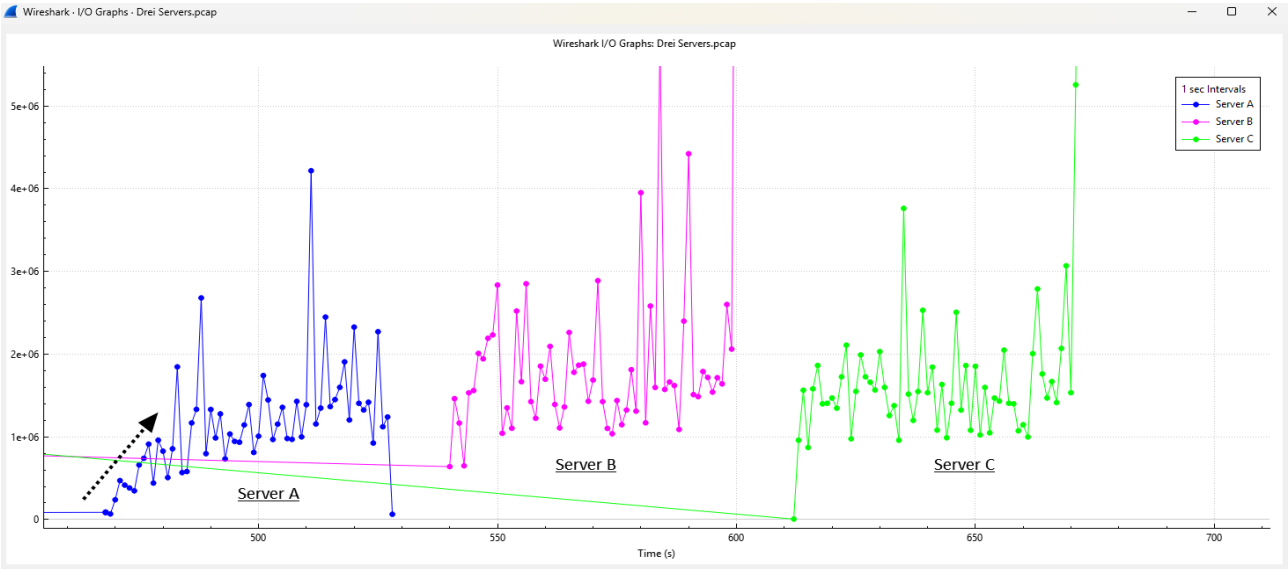
Server A exhibits a slow and gradual increase in Bytes-in-Flight at the beginning of the session. The ramp-up phase is prolonged, and the maximum Bytes-in-Flight remains consistently lower than those observed for Server B and Server C. This behavior indicates a conservative congestion window growth, which limits how quickly the connection can fill the high-capacity link. The lower ceiling suggests either a smaller effective congestion window or stronger sensitivity to congestion signals.

The comparison clearly shows that Server A underutilizes the available bandwidth during the 3rd download session, as evidenced by its slower ramp-up and lower Bytes-in-Flight ceiling. In contrast, Server B and Server C are able to inject more data into the network, resulting in faster convergence to higher throughput levels.

These differences strongly point to transport-layer configuration effects, such as:

- Smaller initial congestion window or conservative congestion control for Server A
- More aggressive congestion window growth for Server B and C
- Possible differences in TCP pacing or receive window limits

Because the sessions are non-overlapping, the observed behavior is not influenced by traffic concurrency and reflects intrinsic differences in TCP dynamics between the servers.

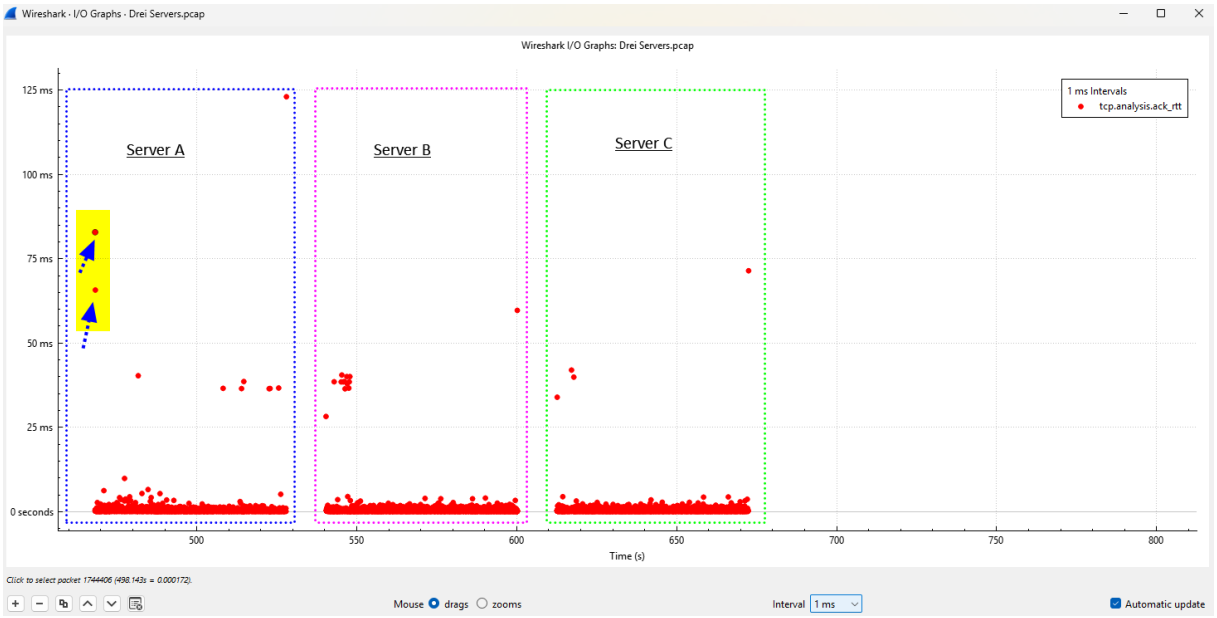


Example: TCP ACK RTT Comparison (3rd Download Session for Server A, B and C)

The figure below shows the TCP ACK Round Trip Time (RTT) measured using tcp.analysis.ack_rtt for the third downlink session of Server A, Server B, and Server C, plotted with a 1ms interval.

Server A shows a significantly higher RTT compared to the other servers. While most RTT samples remain close to the baseline, frequent elevated RTT values are observed in the range of approximately 60- 90ms, with occasional approaching 120ms. These RTT excursions indicate increased path latency or queueing, which directly impacts TCP congestion window growth. The higher and more variable RTT observed for Server A is consistent with its slower Bytes-in-Flight ramp-up and lower sustained throughput.

Server B and C exhibits a lower and more stable RTT distribution, with the majority of samples concentrated at very low values (a few milliseconds), and only occasional outliers around 30–40 ms. This relatively stable RTT allows TCP to increase the congestion window more rapidly, explaining the faster ramp-up and higher Bytes-in-Flight levels observed in the previous analysis.



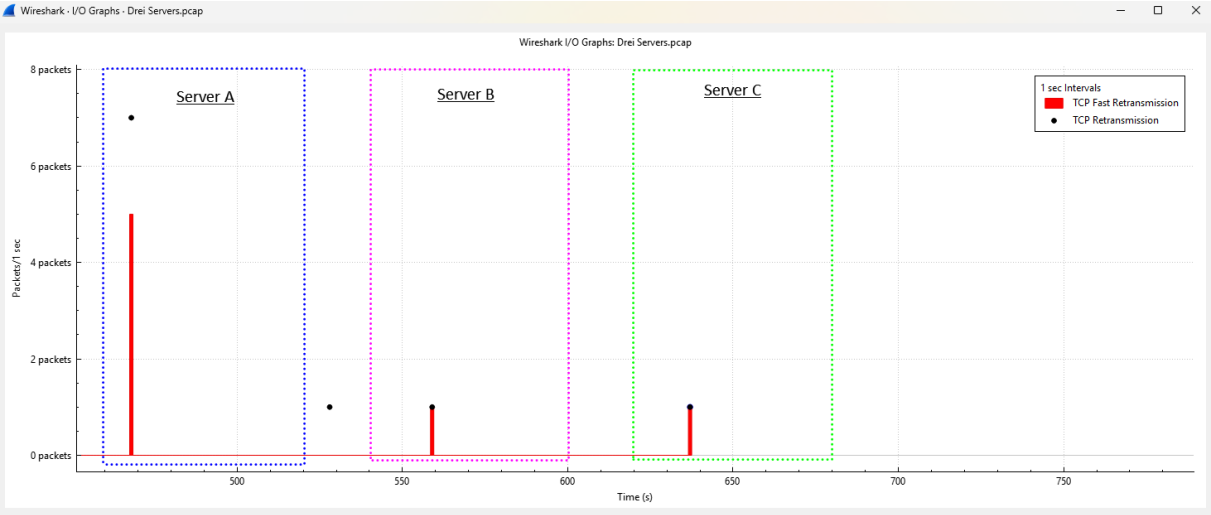
Example: TCP Retransmission Analysis (3rd Download Session for Server A, B and C)

The figure below shows the number of TCP retransmissions per second during the third downlink session for Server A, Server B, and Server C. The graph distinguishes between fast retransmissions (red bars) and retransmissions detected via timeout (black markers).

Server A exhibits the highest retransmission activity among the three servers. Several fast retransmission events are observed, reaching up to approximately 5 packets per second, along with isolated retransmission events. This indicates frequent packet loss or packet reordering along the network path. Such loss events trigger congestion window reductions, which explain the previously observed slower Bytes-in-Flight ramp-up, higher RTT variability, and lower sustained throughput for Server A.

Server B shows very limited retransmission activity, with only isolated single-packet retransmissions observed during the session. The low frequency and small magnitude of retransmissions suggest a stable transport path with minimal packet loss. This behavior is consistent with the faster congestion window growth and higher throughput stability observed earlier.

Server C demonstrates similarly low retransmission rates to Server B. Only occasional, isolated fast retransmissions are present, and no persistent retransmission bursts are observed. This indicates a well-behaved network path with minimal congestion or loss, supporting the high and stable Bytes-in-Flight and low RTT measured for this server.



ROMES 5G DL Throughput Analysis (3rd Download Session for Server A, B and C)

The figure below presents the 5G downlink performance measured by ROMES during three consecutive download sessions, labelled Server A, Server B, and Server C. The analysis includes PDSCH throughput, MCS, resource block (RB) allocation, slot utilization, number of layers / RI, and BLER.

Server A

- Throughput: Server A shows a slow throughput ramp-up, indicated by the gradual increase at the beginning of the session. The downlink throughput exhibits large fluctuations and does not consistently reach the maximum levels observed for the other servers.
- MCS: The average PDSCH MCS for Server A is lower and more unstable, with frequent drops. This indicates that the scheduler is forced to use more robust modulation and coding schemes, reducing spectral efficiency.
- Resource Allocation (RB & Slot Utilization): RB allocation ramps up slowly and shows intermittent drops.
- Slot utilization is irregular, with multiple scheduling gaps. These behaviors indicate inefficient resource utilization, even when radio conditions are sufficient.
- Layers / Rank Indicator: Server A intermittently drops from 4 layers to lower ranks, reducing spatial multiplexing gain and directly impacting throughput.
- BLER: Server A exhibits elevated and unstable PDSCH BLER, with frequent spikes. These BLER events trigger:
  -MCS back-off
  -HARQ retransmissions
  -Reduced effective throughput

<u>Interpretation:</u>

Although radio resources are available, Server A fails to fully exploit them. The slow ramp-up and instability strongly suggest transport-layer limitations (e.g., TCP ramp-up, retransmissions, RTT variability) propagating upward and preventing the MAC scheduler from sustaining high-efficiency transmission.

Server B

- Throughput: Server B reaches high throughput rapidly and maintains a stable plateau throughout the session. Short dips are observed but recovery is fast.
- MCS: The PDSCH MCS remains consistently high, indicating good link adaptation and stable channel quality.
- Resource Allocation: RB allocation is high and stable
- Slot utilization remains near maximum during the active period. This shows that the scheduler can continuously allocate resources without back-pressure.
- Layers / Rank Indicator: Server B maintains 4 layers for most of the session, indicating stable MIMO conditions and effective rank adaptation.
- BLER: BLER remains low and well controlled, with only minor spikes. This confirms efficient HARQ operation and good link quality.

Interpretation:

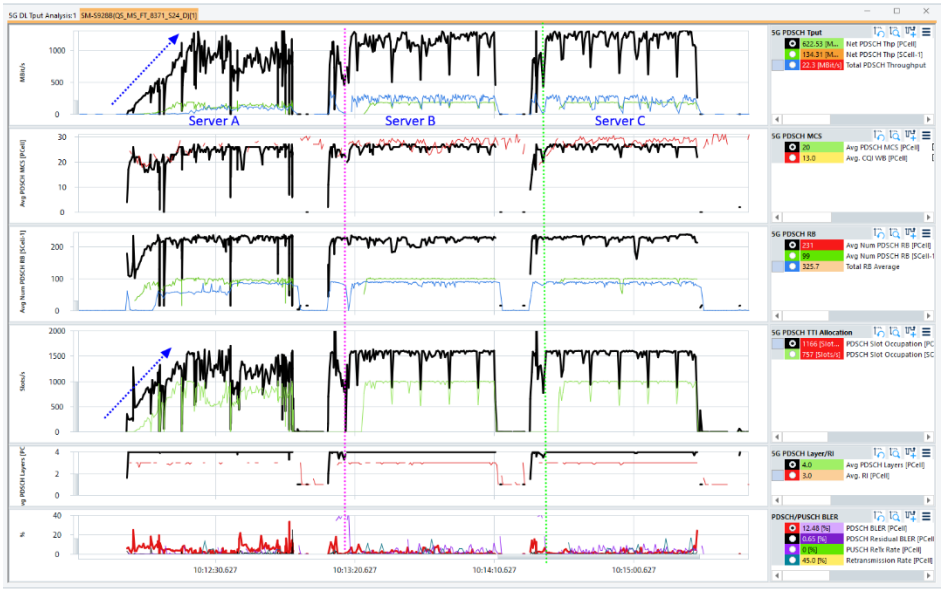
Server B demonstrates balanced and efficient end-to-end performance, where stable transport-layer behavior allows the RAN scheduler to fully utilize radio resources, resulting in sustained high throughput.

Server C

- Throughput: Server C achieves throughput levels comparable to or slightly higher than Server B, with fast ramp-up and high stability.
- MCS: The average MCS is high and stable, with fewer deep drops than Server A.
- Resource Allocation: RB usage is consistently high
- Slot utilization remains close to saturation during the session
- Layers / Rank Indicator: Server C maintains 4-layer transmission almost continuously, indicating strong and stable MIMO conditions.
- BLER: BLER remains very low, with minimal variability, suggesting optimal link adaptation and minimal HARQ overhead.

Interpretation:

Server C represents the best overall performance, combining fast throughput ramp-up, efficient radio resource usage, stable MIMO operation, and low error rates.


Server A
Server B
Server C
10:12:30.627
10:13:20.627
10:14:10.627
10:15:00.627