HyperTransport® CONSORTIUM

Cluster Network Performance

Bandwidth vs. Latency

Definitions

Bandwidth - How much data (number of bits) can be transferred from point (server node) A to point B of the cluster within a given time frame (nS, uS, mS, etc.). Hence, **bandwidth means capacity**.

Latency – The time it takes to complete a data transfer task from point A to point B of the cluster, or the time "delay" between the start of a task/transaction and its completion . Hence

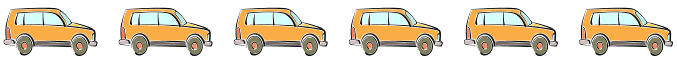
Latency Factor – Any element that increases the time it takes to complete a task/transaction (latency) – i.e. the processing time of interconnect and network transfers, processors, IC controllers and/or any function that stands between data at point A and point B in the cl.

Does bandwidth have a direct influence on performance? Yes, if the available bandwidth is exploited fully. No if otherwise – i.e. with bandwidth margins, increasing bandwidth further does not increase performance.

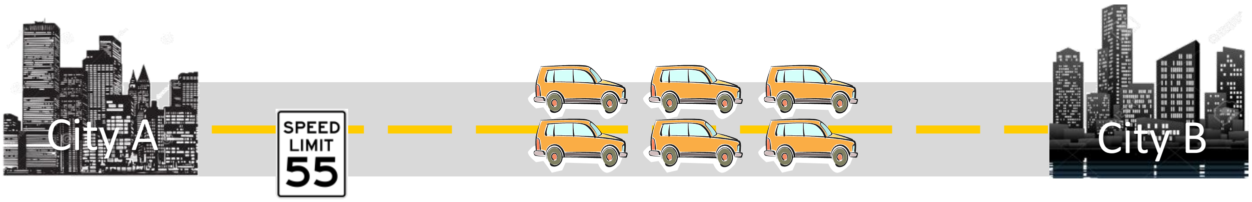
Does latency have a direct influence on performance? Yes. Always. Whether the available bandwidth is exploited fully or not.

Example 1

6 cars need to travel from City A to City B on a 2-lane highway (bandwidth) that has a 55Mph speed limit (latency factor)



The cars can travel 1 per lane, 2 at a time, occupying the whole highway (full bandwidth) at 55 Mph, until all 6 cars have completed the journey (latency)

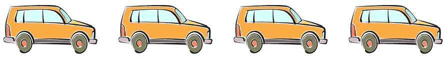


Question: With more lanes (more bandwidth) will it take less for all 6 cars to complete the journey (reduced latency)?

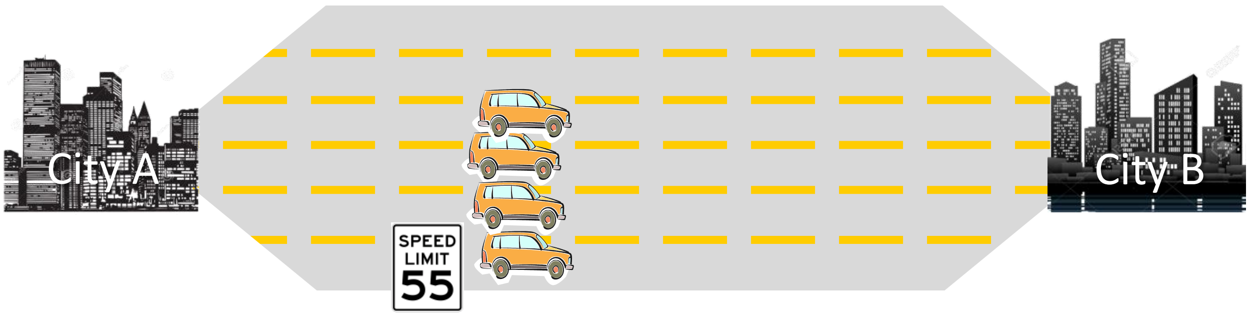
Answer: Yes, because the traveling cars were actually saturating the 2-lane traffic capacity of the highway and could have taken advantage of more lanes.

Example 2

4 cars must travel from City A to City B on a 6-lane highway (bandwidth) that has a 55Mph speed limit (latency)



They can travel 1 per lane, 4 at a time, occupying only 2/3 of the highway at 55 Mph, until all 4 cars have completed the journey



Question: If the highway had more lanes (more bandwidth) would it take less for all 4 cars to complete the journey?

Answer: No, because the highway already had more lanes than the traffic would be able to use (non-saturated bandwidth).

Question: Would an increase in speed limit (latency factor reduction) reduce the journey time (latency) of all examples above?

Answer: Yes, regardless of the number of lanes (bandwidth capacity)

Conclusions

Lower latency Always improves performance. More bandwidth does Only in case of bandwidth saturation - something that in the computing world happens rarely and when it does, temporarily. Therefore:

Lower Latency
is Far More Important than
Higher Bandwidth