On the relation between latency and geographical distance

1 Abstract

With the arrival of edge computing, more and more applications are relying on a tight latency budget and more importantly on shorter latencies. This does not mean that all application components need to be as close as possible, but more that there is a relation between latency and application experience/functionality or similar characteristics.

When data packets are travelling, they do so by the speed-of-light which would kind of suggest geographical distance could be used to calculate the distance. But of course, it is more complicated than that as equipment such as routers are involved in the forwarding of packets. Further, the fibres go in specific directions and aggregate at certain points. Much like how airlines use airports with certain airports acting as hubs, so you must make a detour on the route from A to B (over C).

Some systems use geolocation to try to calculate latencies between application components distributed across a network. This study shows that that method is far from accurate. The relation between distance and latency is good for long distances (more than 4000 km), but the variation for shorter distances is far too big. The variation exhibits more than 4x variation when the distance is shorter.

One should also note that for applications in need for a tight latency budget will require a distance in the shorter scale to achieve the needed characteristics. A distance of 4000 km translates into a latency of around 60ms for comparison, so for latencies shorter than that, we will start to experience variation.

To better understand how latency vary over distance we conducted this study.

2 Collection of basic data

Firstly, we identified 26 different locations in Europe and the US. They were selected to give a spread in their geographical distance so that we would get some locations close to each other as well as being both on larger distances both in Europe as well as in the US.

To get data on the latency between these locations we used ping statistics from WonderNetwork (https://wondernetwork.com/pings) and calculated the geographical distance between them by using https://www.distancecalculator.net/

3 Europe (and some US locations)

Our first data set is 17 locations (14 in Europe and 3 in the US). The US locations is there to provide extremely long distances. So, if you plot latency over distance, you get this diagram:

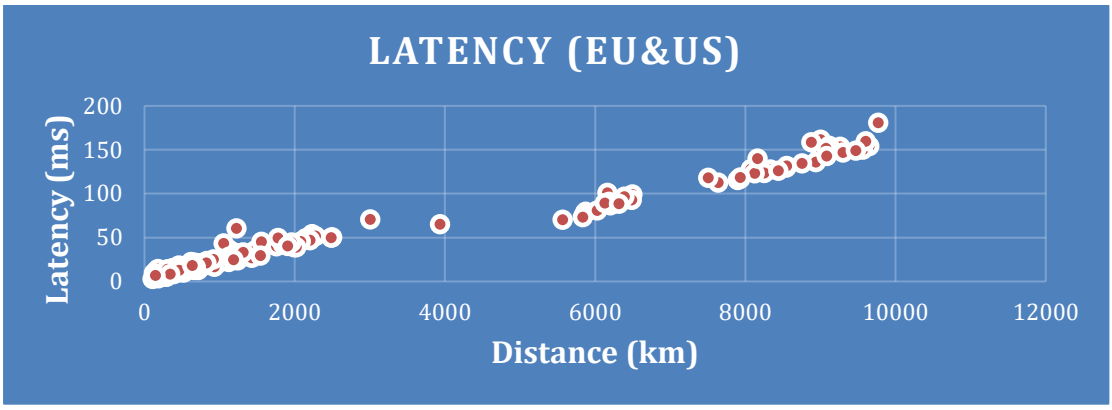

Figure 1:Latency over Distance

This shows quite a nice relation with some smaller outliers.

Let's now calculate the latency as milliseconds per km and plot that over the distance. This will illustrate the variation.

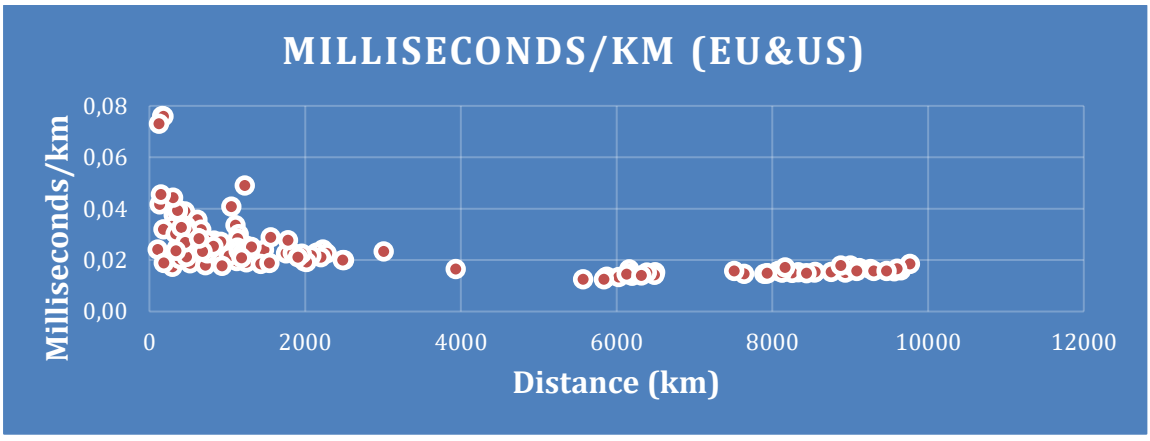

Figure 2: ms/km over Distance

Here we clearly see a huge variation for smaller distances, whilst the longer distances the latency in terms of travel time is quite constant at somewhere around 0,015 ms/km.

In simple terms, we see that the detours that packets must take becomes a larger portion of the total latency for shorter distances. This is intuitive. We also have some outliers where that effect was particularly bad. The three most notable ones are: Falkenstein to Munich, Dusseldorf to Frankfurt and Antwerp to Dusseldorf all having a travel time above 0,07 ms/km.

The measures having the longest distance represents Europe to US routes that obviously travel by an Atlantic sub-sea cable for a long part of the distance, thereby not varying very much.

So, if we then drill down into the part of the diagram that has the shorter latencies, we get the following plot.

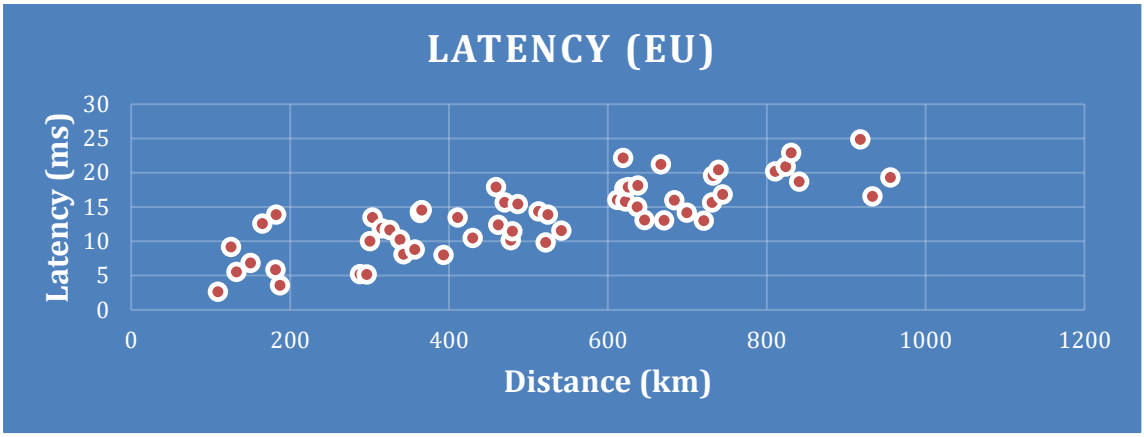

Figure 3:Latency over Distance, Europe only

Here, the data points are the subset of the previous having a geographical distance less than 1000 km. This means they are all withing Europe. So now we see the variation clearer.

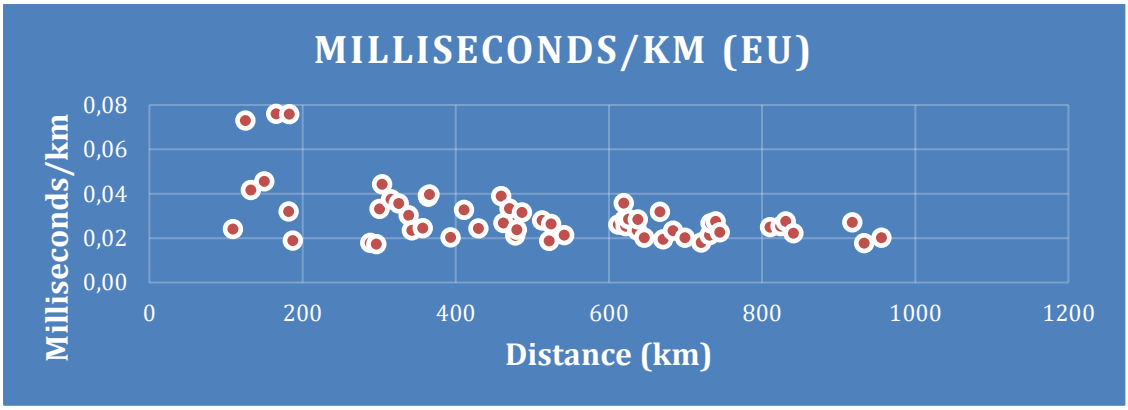

Figure 4:ms/km over Distance, Europe only

We see now quite clearly that the variation is bigger the shorter the distance and converge at longer distances.

So, the conclusion is that for latencies lower than 20 ms (corresponding to geographical distance of 1000 km) it is not possible to use geographical distance as the way to calculate the latency. The error is 4x or more and predicting optimistic values.

4 US Locations

To check if this is true also for the US, we took the three US locations from the first data set and added another 11 US Locations. Also, this time with a spread in locations that are close (on the east coast) and locations that are farther away (on the west coast or south).

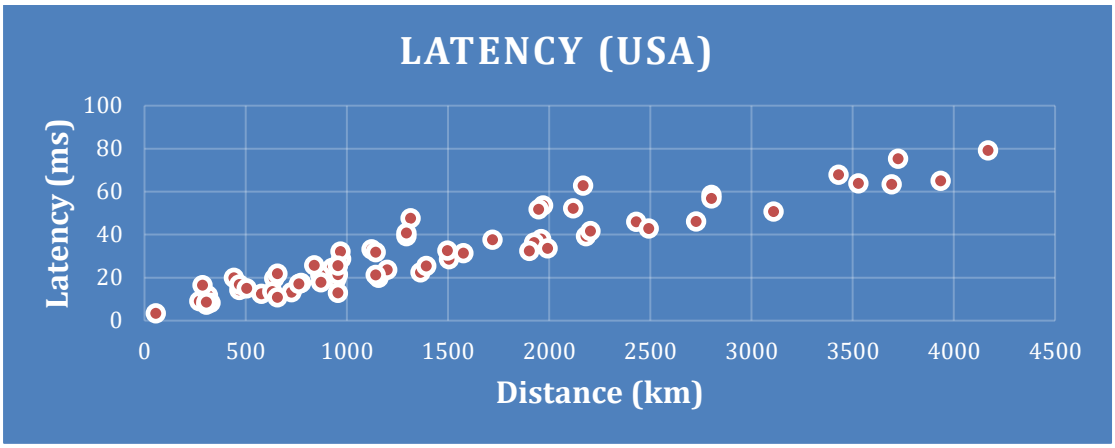

Figure 5:Latency over Distance, US only

Data is consistent with Europe, and one could see a variation that seem to be less at shorter distances.

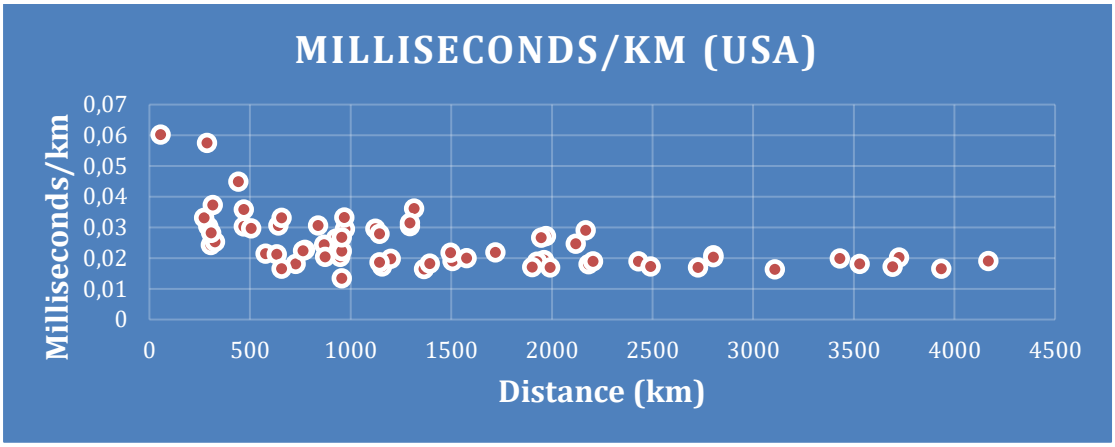

Figure 6:ms/km over Distance, US only

But when calculating the ms/km metric we again get a consistent view. The difference for the US seems to be that the variation is bigger for longer distances. So, while Europe had converged the figures somewhere around 1000 km, the US seems to do so at around 2500 km.

This can be explained perhaps by how the locations were selected and the larger geographical spread of the US population causing a less dense network.

It is only 6 US mainland states that has a population density higher than that of the whole of Germany, so depending on how the locations are chosen, the picture would look different.

The overall conclusion remains. I.e. that for emerging low-latency services, geographical distance is not a good method by which latency can be calculated.