350 million people within 30 ms – is Sweden the gateway to a sustainable future?

This White Paper has been produced in a collaboration between Invest Stockholm, Telia Carrier and Stockholm Data Parks. It was written in the third quarter of 2019 and the principal authors were **Mattias Fridström** of Telia Carrier and **Johan Börje** of Stockholm Data Parks/Stockholm Exergi.

Table of contents

Introduction

The choice of a data center's location is determined by multiple factors, which vary with the specific needs of different market segments. In this report we address one of the main site location criteria – network latency, also referred to as lag or delay.

Network latency is ultimately a question of service quality – if it gets too high, a user's perception of quality can be adversely affected. The degree to which service degradation is perceived varies with the type of application that a user is running.

But beyond that, optical technology and low latency transmission have a more profound impact on the modern world – a world searching for ways to secure a sustainable future. By enabling rapid transmission of data, fiber optic networks can effectively be used to manage energy resources more efficiently and re-distribute power production and consumption across a broader geographical area.

This report explores the vast geographical end-user market and breadth of applications that could be served from a Stockholm data center, without tangible effects on a user's perception of service quality but with a positive effect on the environment and operational cost.

Executive summary

With green energy, the lowest electricity prices in the EU and an opportunity to receive payment for heat recovery – as well as a vibrant business environment with a large appetite for digital services, there are many compelling reasons to locate a data center in Sweden's capital, Stockholm.

A key finding of this white paper is that, with Europe's population concentrated towards the north, Stockholm is located within 30 milliseconds round trip delay from a potential digital market of more than 350 million end-users. The sheer size of this market is enabled by Telia Carrier's extensive, multi-path fiber network connecting the major cities of Europe with minimal delay.

While some applications depend on ultra-low latency and therefore close physical proximity to server resources, the majority of applications can be provided to northern European end-users from Stockholm, with a first-rate customer experience.

Data center relocation to Sweden, and Stockholm in particular, presents a compelling opportunity for an energy-intensive industry to source cheap electricity with a low carbon footprint, whilst operating within acceptable latency limits. It is possible to implement a site location strategy for Northern Europe, whereby a significant share of users' applications are hosted remotely in Stockholm - with competitive OPEX, real estate costs and mitigation of negative environmental effects through the use of renewable electricity sources and payment for reuse of excess heat. This approach can deliver significant annual cost savings for data center operators and their customers.

At the same time, much-needed production and grid capacity can be freed-up in in other European centers, where the opportunity for expansion is sometimes more expensive and constrained.

This makes Stockholm an ideal location for hosting and processing data that is consumed throughout the Nordic countries, Baltic states, western Russia and northern Europe in general.

Addressable Market

Europe's population is concentrated to the North.

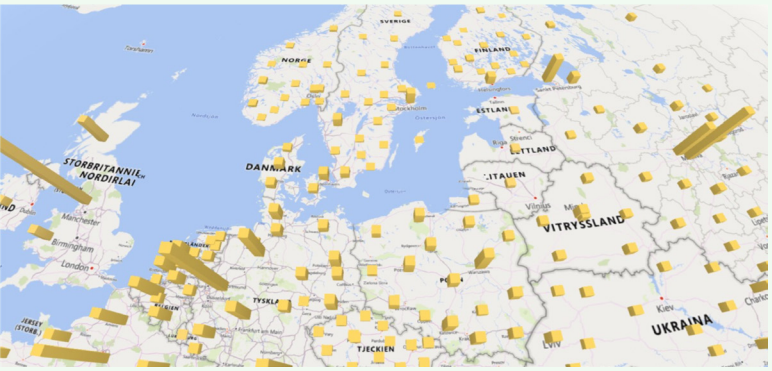


Figure 1 – Northern Europe's population centers. The yellow bars represent population size, either on a regional or national level (to increase readability). Source: https://www.citypopulation.de/, with recent estimates during May 2019, mapped in MS Excel and Bing Maps.

The map in Figure 1 covers countries with a combined population in excess of 437 million inhabitants, including the Northwestern and Moscow (Central) regions of Russia.[1] To what extent the citizens and companies in these countries can be considered part of an addressable data center market from Stockholm depends to a great extent upon the latency between Stockholm and the target geography. A critical component of that analysis will be the existence of fiber carriers' Points-of-Presence (PoPs) in key metropolitan locations. As we will see later on, the vast majority of the highly populated areas in Figure 1 can be reached with minimal delay thanks to Telia Carrier's extensive network footprint.

Country	Population
Germany	82,792,351
France	64,812,052
Great Britain	55,619,430
Ukraine	44,736,804
Russia CFD	39,311,413
Poland	38,433,558
Netherlands	17,282,753
Russia NWFD	13,952,003
Belgium	11,376,070
Czech Republic	10,610,055
Sweden	10,230,185
Belarus	9,475,174
Denmark	5,806,081
Finland	5,517,919
Scotland	5,424,800
Norway	5,328,212
Ireland	4,761,865
Wales (Cymru)	3,125,165
Lithuania	2,793,986
Latvia	1,934,379
Northern Ireland	1,870,834
Estonia	1,319,133
Luxembourg	613,894
Iceland	356,991
Sum	**437,485,107**

Figure 2 – Population size of countries included in Figure 1.

[1] According to Wikipedia, Europe's total population is around 741 million, of which 110 million are Russians to the west of the Ural Mountains (77% of the total Russian population). Roughly half of these, or 53 million people are included in the 437 million figure above.

Energy production in Sweden

Sweden's electrical power comes primarily from hydro and nuclear sources. Together, they account for 80 per cent of all electricity production in the country. The remainder comes largely from wind and co-generation of heat and power.

In 2017, total electricity production amounted to 159 TWh. Hydropower accounted for 64 TWh, which represents 40 per cent of Sweden's total electricity production capacity. During a typical year, approximately 65 TWh of electricity is produced this way, but fluctuating precipitation can cause this to deviate by around 15 TWh.

Nuclear power accounted for 63 TWh in 2017, corresponding to a 40 percent share of total production capacity. Wind power accounted for 11 percent and co-generation 15 percent. In 2019, wind energy production is expected to reach 20 TWh and current estimates indicate that wind power will continue to grow rapidly.

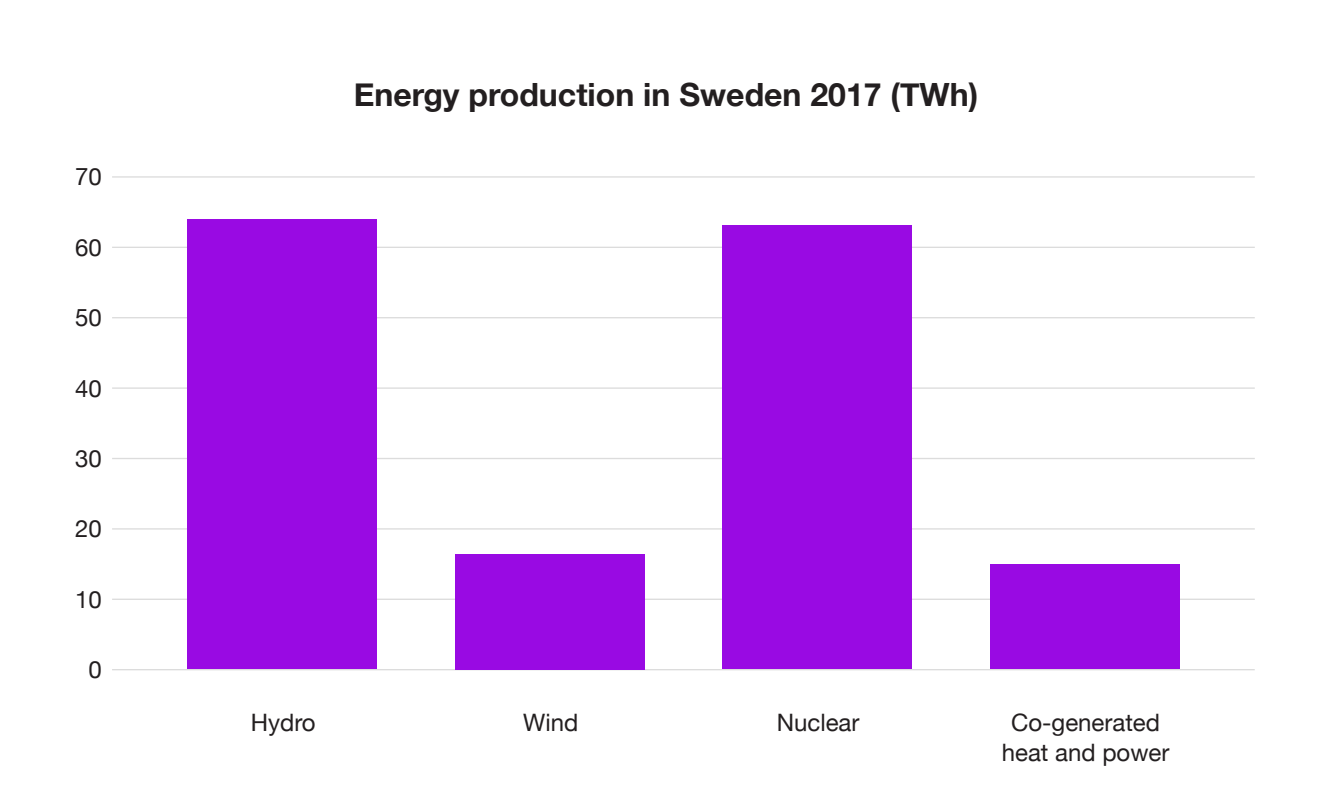


Figure 3 – Energy Production in Sweden. Source: SCB

Overall, Sweden's high proportion of hydro, wind and nuclear power make it a very low producer of hydrocarbon-based electricity, and it therefore has a competitive green electricity footprint when compared with other European countries. At the same time, Sweden has some of the lowest energy prices in Europe.

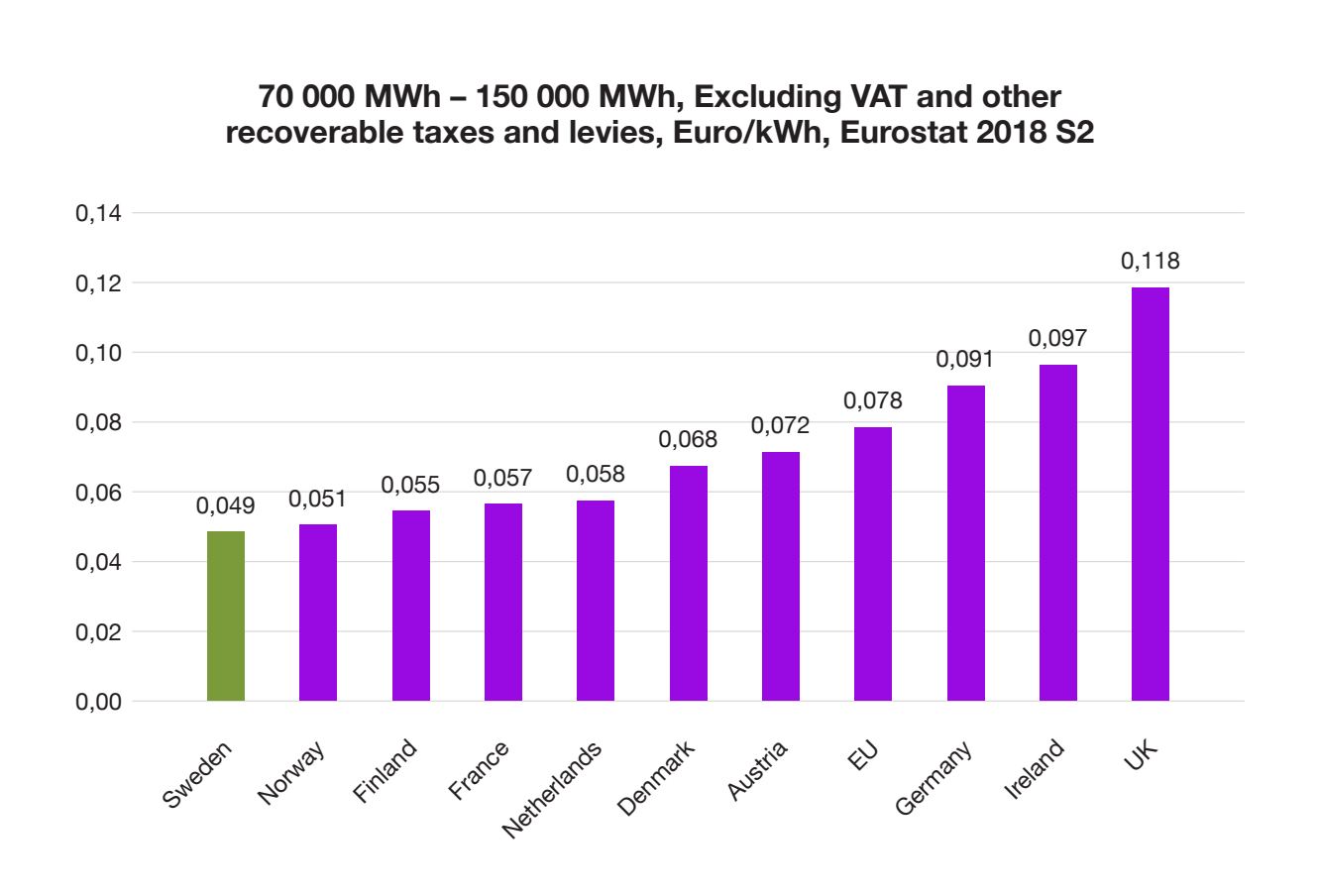


Figure 4 – Energy prices in Europe. Source: Eurostat 2018. Please note: Whilst the price is applicable for data centers in Sweden, in other countries data centers may be liable to additional fees. In Germany, for example, data centers are obliged to pay the so-called 'EEG-Umlage', which results in a kWh price of approximately 16 euro cents.

Telia Carrier's network

Telia Carrier is a provider of global backbone connectivity that enables operators, content providers and enterprises to deliver critical services, without the need to worry about how much bandwidth they require, to whom they need to connect, or where their data is consumed.

Telia Carrier's story stretches back to 1993 when AS1299, their global Internet backbone, was established.[2] In terms of Internet connectivity, AS 1299 has sustained the #1Global IP Backbone network position, since 2017.[3] Its global IP Transit customer base currently accounts for 58% of global Internet routes.[4]

Spanning 65,000 km of fiber assets throughout Europe and North America, Telia Carrier's global network connects more than 300 Points-of-Presence worldwide. The network is engineered to meet the traffic demands of both public Internet and private networks, and is currently growing by around 25-30% per year.

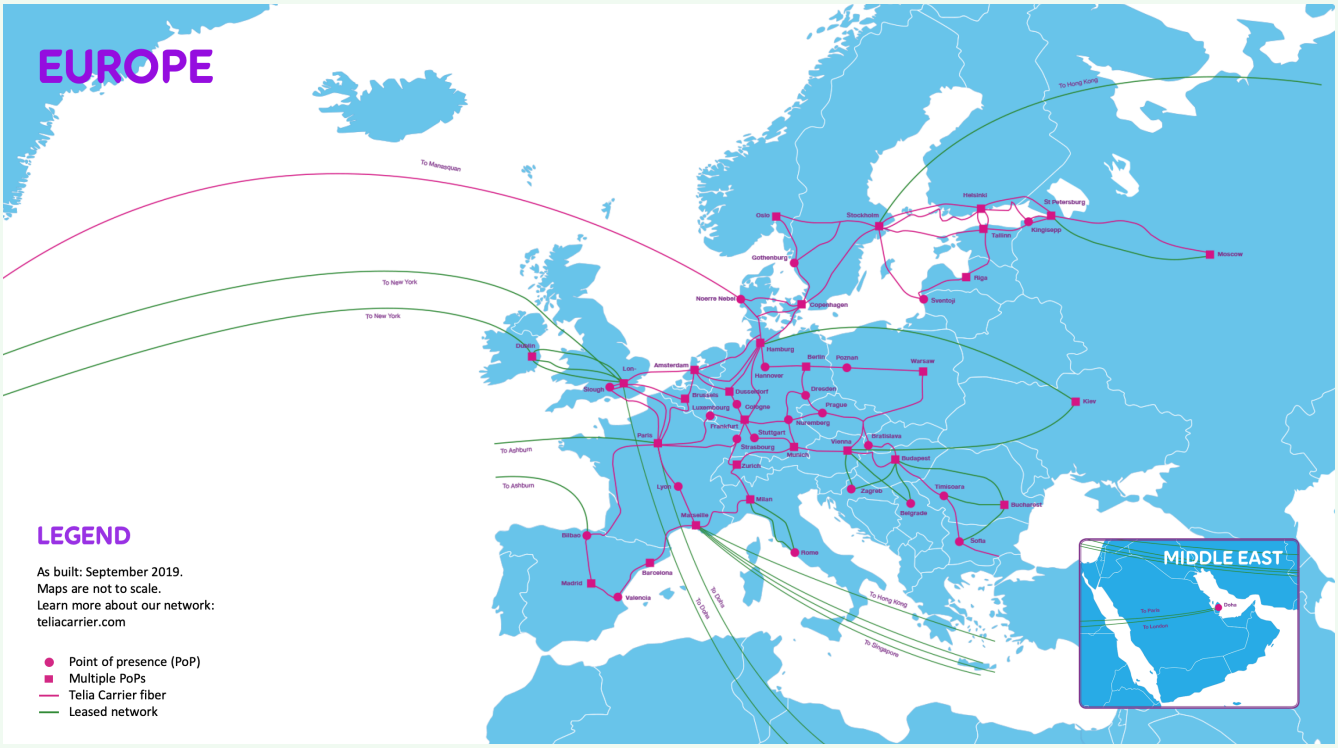


Figure 5 – Telia Carrier's international core network. Please note: This does not include national network infrastructure.

[2] AS is Autonomous System, which is the name for the independent networks which together make up the Internet.
[3] DYN IP Transit Intelligence, Backbone Rankings, May 2019
[4] CAIDA AS rankings, July 2019

The Nordic region is particularly well connected and all countries in the region were early adopters of fiber technology, more than three decades ago. This is especially true for the Stockholm region which is arguably one of the best-connected in the world. Here, an extensive network of fiber optic infrastructure connects locations in all geographical directions: to the north of Sweden – all the way into the Arctic Circle, towards Oslo in the west, eastwards with cables to Helsinki, Tallinn, Riga and Vilnius via the Lithuanian coast then St. Petersburg and onwards into Russia. And finally, towards the south with multiple paths towards Copenhagen, Western Europe, the US, Asia and beyond. In addition to Telia Carrier, there are twelve other international carriers offering services in Stockholm.

Stockholm is also amongst the cities in the world with the densest metropolitan fiber footprint. STOKAB – a company owned by the city of Stockholm and part of Stockholm Data Parks – currently connects 24,000 facilities in and around the city. Furthermore, operators like Telia (the former incumbent), Tele2, IP-Only and others have also built their own fiber networks, contributing to a healthy and competitive marketplace.

All-in-all, this makes Stockholm stand out in terms of both local and international connectivity. A quick review of the four major Capital cities in the Nordics also shows that significantly more traffic is exchanged in Stockholm than Oslo, Helsinki or Copenhagen. This is mainly due to Stockholm's central location in the Nordics, where a larger market can be served from a single point and also because its extensive city network reaches almost every building with fiber.

Stockholm's IXP – Netnod

The vast majority of global Internet backbone traffic transits a limited number of private Internet backbone networks, either as commercial IP transit traffic across them, or as so-called 'peering' traffic between them.

On a local and regional level, Internet exchange points exist. An Internet exchange point (IX or IXP) is the physical infrastructure through which Internet service providers (ISPs) and content delivery networks (CDNs) exchange regional Internet traffic between their respective networks (autonomous systems).[5]

The larger a city's IXP, the easier and cheaper it is for service providers, in particular, to exchange traffic with content networks and operators in the local market and even into adjacent regions beyond. In addition to the connectivity offered by Telia Carrier and other international Carriers, Netnod operates the largest IXP in the Nordics and provides connectivity throughout the region.

Netnod is present in several Stockholm data centers and has the possibility to reach other new data centers indirectly. At these locations, Netnod presents an opportunity to peer with (or connect to) some of the largest transit providers, telcos and CDNs in the region. Currently, Netnod in Stockholm provides connectivity to 168 unique ASes.[6]

[5] https://en.wikipedia.org/wiki/Internet_exchange_point
[6] https://www.netnod.se/ix/networks

Latency to northern Europe

One of the most widely-discussed topics in the telecom world is the time it takes for data traffic to reach a particular destination and come back again. This is referred to as Round Trip Delay (RTD) and quite simply, it measures (in milliseconds) the time it takes for traffic to transit the network from its point of origin to a specific destination and back.

Often, this is also used to define the time taken, or latency, of traffic as it propagates a particular fiber stretch. Low latency means that traffic travels quickly along a network path and higher latency naturally increases the time for data packets to travel between source, destination and back again.

The reason for measuring the round trip, rather than a single direction, is that a single source can measure and objectively compare different destinations with a good degree of accuracy. In practical terms, many applications are also dependent on getting an acknowledgement from the receiver, even though new packets of data can be sent before preceding packets are acknowledged (e.g. HTTP/2 multiplexing and server push over TCP). That said, real time applications typically don't rely on acknowledgements and make use of User Datagram Protocol (UDP) packets.

Since the vast majority of all traffic runs through fiber cables, latency is ultimately dependent upon the physical distance between the start and end points of the traffic. Optical fiber is made of glass and the transmission of traffic through fiber-optic cables is essentially broken down into a never-ending stream of photons that travel along them. A group of photons means a "one" and an empty gap between them means a "zero". In turn, these ones and zeros are re-combined at the destination to form the "message" being sent.

Since photons are used to carry data, the speed of light will be a limiting factor in defining the minimum possible latency between two destinations. The speed of light at normal air pressure will always be around 299,792,459 m/s (often simplified as $3x10^8$ m/s). This number could be used to calculate latency, but only if conditions are absolutely perfect. In reality, cables are bent and apart from degradation of the optical quality over time, cables are continuously repaired and connectors added, reducing the ability of the fiber to efficiently transmit light over time. Ultimately, this will impede the photons in the cable and slow things down. Therefore, a rule-of-thumb used in calculations is that the speed of traffic is 2/3 of the speed of light. $2x10^8$ m/s is therefore often used to calculate the latency between two locations.

If fiber distance affects latency, then of course the routing of the cable itself has a tangible impact on network performance. Cables are seldomly routed in a perfectly straight line between two points. Most of the time, they are built in conjunction with other infrastructure projects , often following roads, electricity or gas lines, rivers or railways.

Since cable stretches are expensive to deploy, utilization needs to be maximized by passing as many potential interconnection points as possible between major cities, to enable local traffic 'drops'. This leads to diversions and additional cable length, increasing latency.

Round Trip Delay or latency is simply ascertained by looking at the time it takes for a test packet to be sent and received from one network node to another network destination and back again. It could be from an already installed router in the network or from a test box adding traffic to the network. One very common way of measuring round-trip delay is to use a service called Ping.

Ping uses the Internet Control Message Protocol (ICMP) function “echo request” which simply requests the recipient to send the received packet as an immediate response. Because ICMP is only intended for diagnostic or control purposes it does not provide an exact measurement but rather, a fair indication of performance.

For the purposes of this white paper, Telia Carrier carried out detailed measurements with professional equipment to document the RTD from Stockholm to its Points-of-Presence in selected cities throughout northern and central Europe. The results are summarized in figure 6.

These figures are measured across a dedicated fiber path in Telia Carrier’s commercial network where no risk for unpredictable delay due to buffering in intermediate third-party routers will occur.

It should be noted that for most real-time applications running the UDP protocol, the relevant latency figure is half the RTD number, i.e. the one-way propagation time.

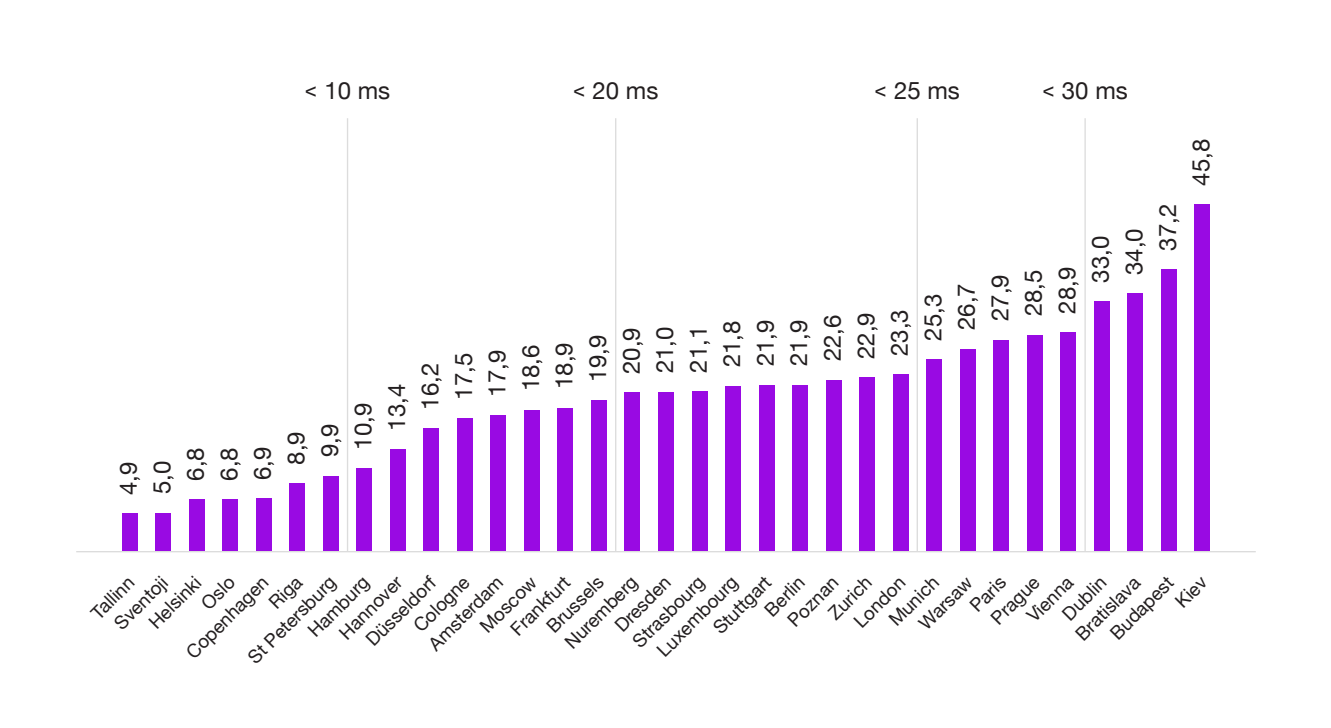


Figure 6 – RTD in milliseconds from Stockholm to a selection of Telia Carrier PoPs.

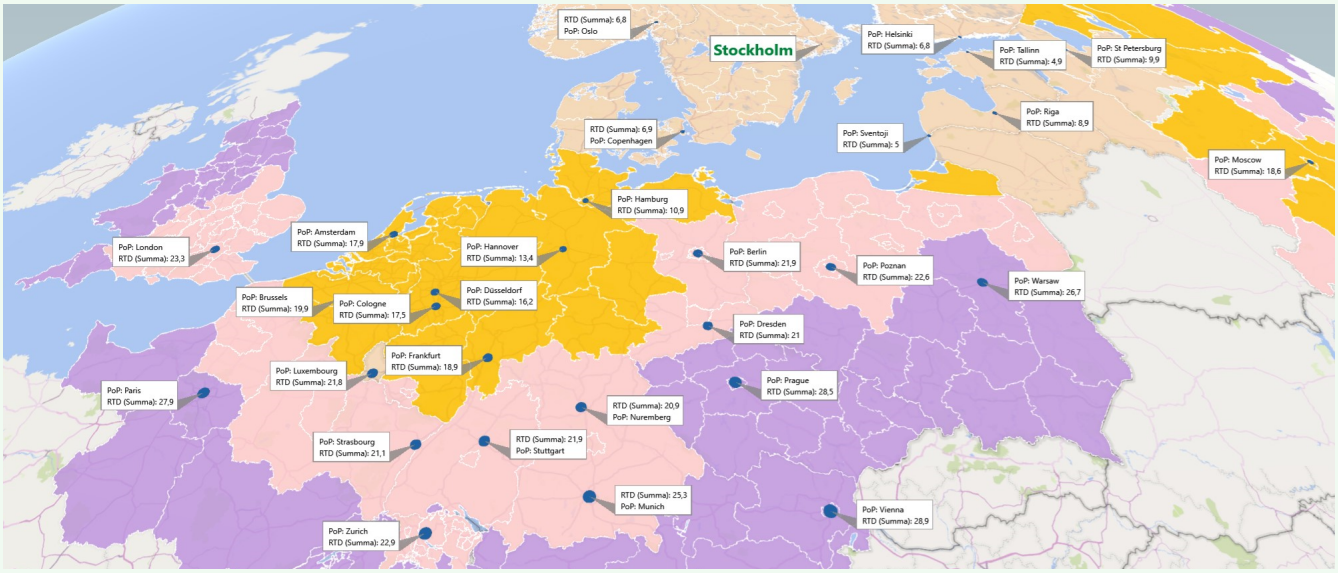


Figure 7 – PoPs, RTDs and reachable geographical areas.

Potential users within fixed RTD ranges

With the population data and Telia Carrier measurements, it is possible to create an estimate of the addressable market within different ranges of RTD.

In figure 7 (see previous page), the PoPs and their RTD have been documented. Color coding has then been applied to distinct geographical areas where a 10, 20, 25 and 30 milliseconds RTD from Stockholm can be reasonably achieved.

By calculating the total population size across these areas, it was found that more than 350 million potential end-users are within a 30 millisecond round trip delay margin from Stockholm.

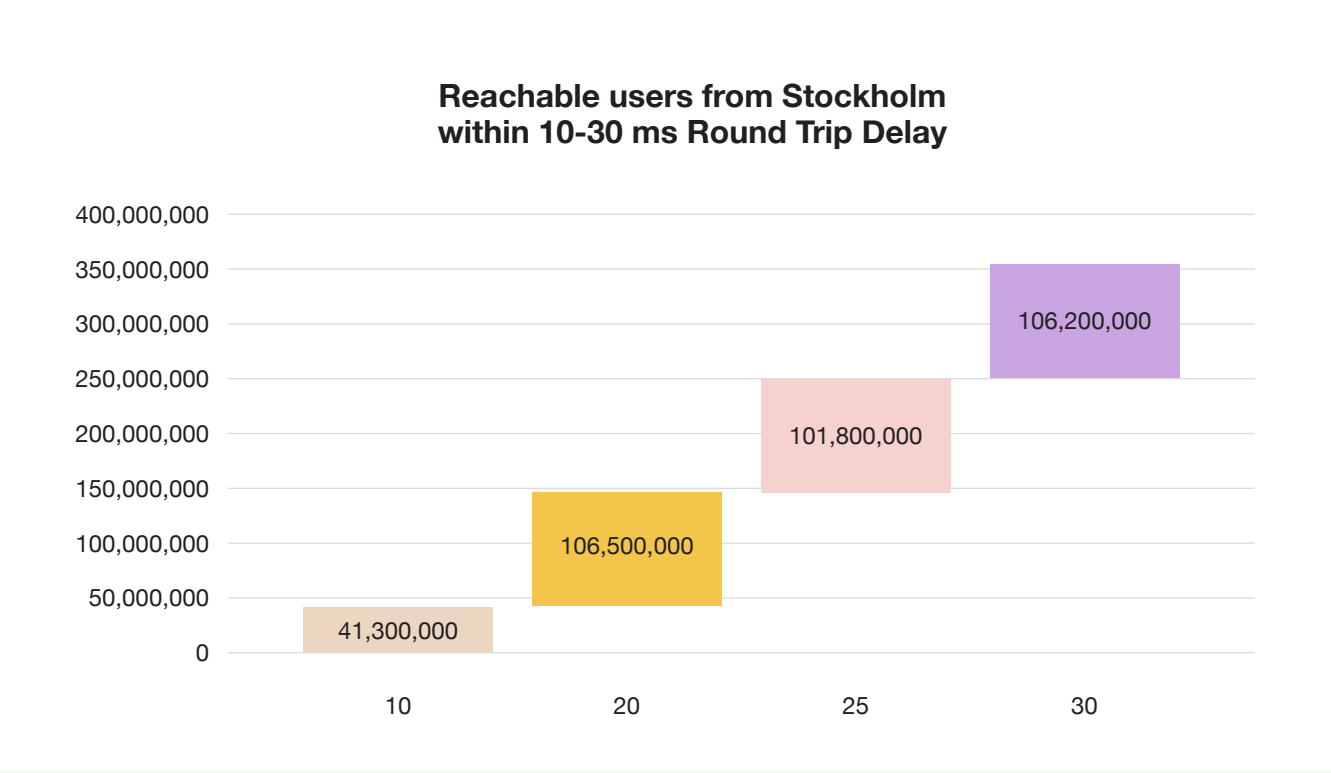


Figure 8 – Potential market from Stockholm in RTD-intervals in milliseconds.

Transmission costs

Enterprises, content providers and operators needing to connect remote sites, cloud resources or such like, will generally look to buy transmission services from international carriers.

German providers for example, buying smart colocation in the Stockholm area, would need to secure traffic to and from Germany. This could be in the form of raw wavelength capacity or lower dedicated bandwidth across an Ethernet interface. In both cases, pricing is generally dictated by the level of traffic commitment, and to a much lesser extent the physical distance between connected sites.

A typical example would be €500 per month for 10 Gb/s of leased wavelength capacity between Frankfurt and Stockholm. This secures a single link of 100% dedicated capacity between the customer location in Frankfurt and the PoP in Stockholm. Where equipment is located in a third-party data center, there are additional costs to be considered, including a so-called ‘cross-connect’ fee from the Colocation service provider to facilitate access to Telia Carrier’s equipment.

Most customers need extra security and an additional back-up path for their traffic. Therefore, it may well be necessary to add another 10 Gb/s wavelength, at a cost of €500/month. This would then ensure a completely separate, additional path through the network between the two locations. This could be used to keep traffic flowing in the event of a network outage on the primary route.

Dual entries into the buildings on both sides and completely diverse cable systems all the way between Frankfurt and Stockholm would ensure that a single fault would not cause both routes to fail. As Telia Carrier’s network is well established across Europe, the RTD time would be comparable across the two routes, allowing the continued use of delay-sensitive applications.

In this example, the cost of transmission between Stockholm and a site in Northern Europe would not be a significant factor when selecting a site or deciding where to store data or processing resources. It should be noted however, that these prices are available for the Stockholm area and other, more remote locations would have a higher price tag.

Applications and latency

Whether the latency added by network overhead is important or not will depend upon the nature of an application and the user's perception of performance. The total application latency will, in addition to the network latency, be the sum of delay contributed by the hardware, software and application implementation of the user's system, as well as the load on any external servers that the application relies on.

For the user, it is the total quality of experience that counts. As can be seen from Table 1, quality of experience is affected by a number of factors that vary by application.

As the table shows, low latency is important for some applications, but far from all. Interestingly, jitter (the variation of delay) is a more important factor for time-sensitive applications. The amount of jitter can typically be managed by a well-engineered network with limited congestion and end-user device buffering.

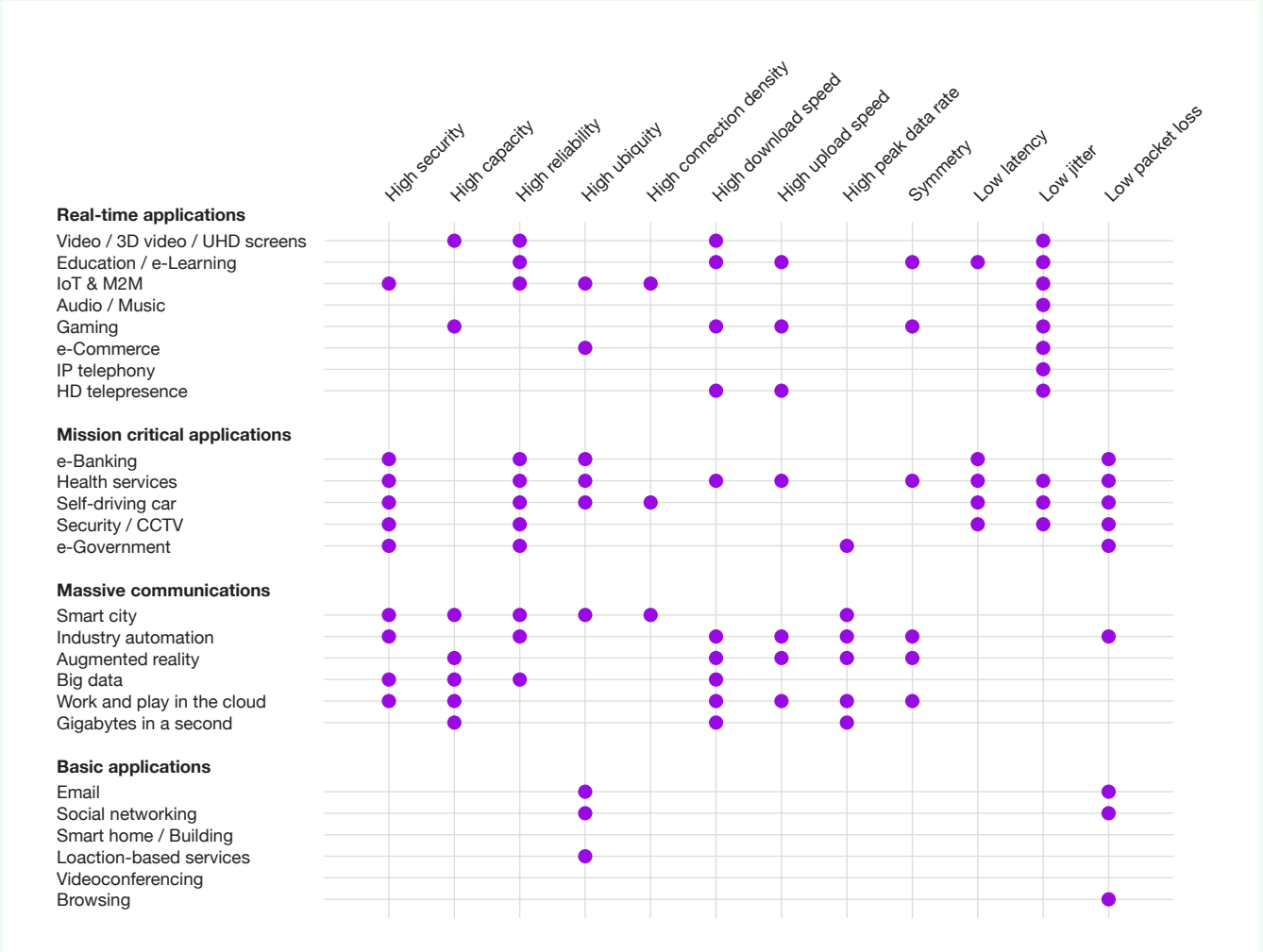


Table 1 – Components of quality for different applications. Source: Connectivity for a Competitive Digital Single Market – Towards a European Gigabit Society, Commission Staff Working Document, SWD (2016) 300 final

As we have already seen, different applications will have different requirements. Also, different users will perceive latency differently. Below is a summary of typical latency requirements for various applications.

IP-telephony works well with a total one-way (as opposed to round trip) latency of 150 ms and that is also the typical target for **video conferencing**, in line with the ITU G.114 standard. When it comes to relative latency for video, the human brain is quite forgiving when a video sequence is out of sync with the soundtrack (lip sync). In fact, the audio can arrive up to 45 ms before and 125 ms after the video without distracting the viewer.[7]

For on-demand **video streaming**, latency is not a critical requirement and delay is generally managed by initial buffering during the initial seconds of a stream. However, when it comes to live event streams, latency can create problems when, for instance, the result of a match reaches the viewer earlier on a different communication channel.[8]

For Citrix and **thin client business applications**, 150 ms is the typical latency limit.[9] [10] For many of these and other services, proprietary protocols are developed for endpoints in order to further reduce latency.

For **web browsing**, requirements can differ considerably depending on the type and purpose of a website. 2,000 – 4,000 ms seems to be considered a tolerable range for webpages to load.[11] To complete the process, the end-user device has to send multiple requests and receive acknowledgements, since a webpage typically consists of several parts. At the heart of these requests are the TCP and HTTP protocols. Historically, this work was to a large degree carried out sequentially, adding unnecessary web browsing delay.

Since 2016, most browsers support the HTTP/2 protocol, on top of TCP. With the introduction of HTTP/2 (RFC 7540), browsing has become much faster thanks to a range of smart techniques that perform parallel as well as proactive data transfer of website content. Work is currently ongoing with HTTP/3 which is expected to further reduce latency in web browsing by using the UDP protocol and moving session integrity into the HTTP-layer.[12] The new protocol, which in its proprietary form is already used by Google with Chrome, is expected to become standard during 2019.[13]

For real-time applications like IP-telephony such as Skype and FaceTime, a different protocol, UDP, is typically used. This doesn't rely on requests and acknowledgements like TCP for the voice packets. To compensate for any resulting bit errors, these applications have other, less intrusive mechanisms to ensure acceptable quality.

It follows therefore, that with UDP applications, the RTD figures presented in previous chapters provide an excessive impression of latency. For UDP, a 'one-way protocol', the network latency for applications will generally be half that of the RTD specified, since no acknowledgements are required. For TCP applications, the delay contributed by the network will impact the application latency, but the extent of this depends on several factors such as client buffer size, protocol and application implementation as well as link quality.

[7] https://support.biamp.com/Tesira/Video/Video_and_network_latency

[8] https://www.theoplayer.com/blog/the-importance-of-low-latency-in-video-streaming

[9] https://www.pubnub.com/blog/how-fast-is-realtime-human-perception-and-technology/

[10] Quantifying Interactive User Experience on Thin Clients, http://isr.cmu.edu/doc/tolia06-ieee.pdf

[11] https://ux.stackexchange.com/questions/58163/acceptable-waiting-time-for-users-in-time-sensitive-actions, http://www.websiteoptimization.com/speed/tweak/psychology-web-performance/

[12] https://en.wikipedia.org/wiki/QUIC

[13] https://kinsta.com/blog/http3/

For **e-commerce sites**, excess latency can have a detrimental economic impact. Amazon reported in 2006 that, according to their calculations, a 100 ms latency increase could reduce sales by 1% (without indicating a specific reference value).[14] In 2009, Google found that by increasing search latency from 100 to 400 ms, the daily number of searches falls between 0.2 to 0.6% (without stating a specific reference latency value).[15] As indicated above, with the implementation of HTTP/3, it looks like future web applications will be running HTTP/3 and that average latency times will reduce as a result.

When it comes to **online or cloud gaming**, an end-to-end, one-way network latency of 80 ms seems to be the upper limit, depending on which type of game is being played.[16] For the most time-sensitive gaming applications, user experience seems to be affected at a total latency threshold of 50 ms, of which 20 ms could be attributed to processing overhead.[17]

In these, and other time-sensitive applications, user experience can be improved by the deployment of edge servers closer to the user (over and above general application optimization, for example the distribution of application and processing routines between the cloud and the client). Cloud gaming often uses both TCP and UDP to combine quality with minimal latency. The deployment of edge servers can also reduce network related content distribution costs, in particular for "heavy" content such as streamed video.

For **Virtual and Augmented Reality**, the latency threshold is somewhere between 19-50 ms.[18] This also applies to **certain IoT-applications**. An example would be autonomous vehicles where a commonly cited acceptable maximum latency level is 20 ms.[19]

For financial **trading**, with machine-to-machine communication, where a single millisecond could have detrimental financial consequences, servers must be colocated at the financial exchange itself. However, all trading does of course not always take place within the same exchange, and the main challenge is to reduce long-distance network delay.

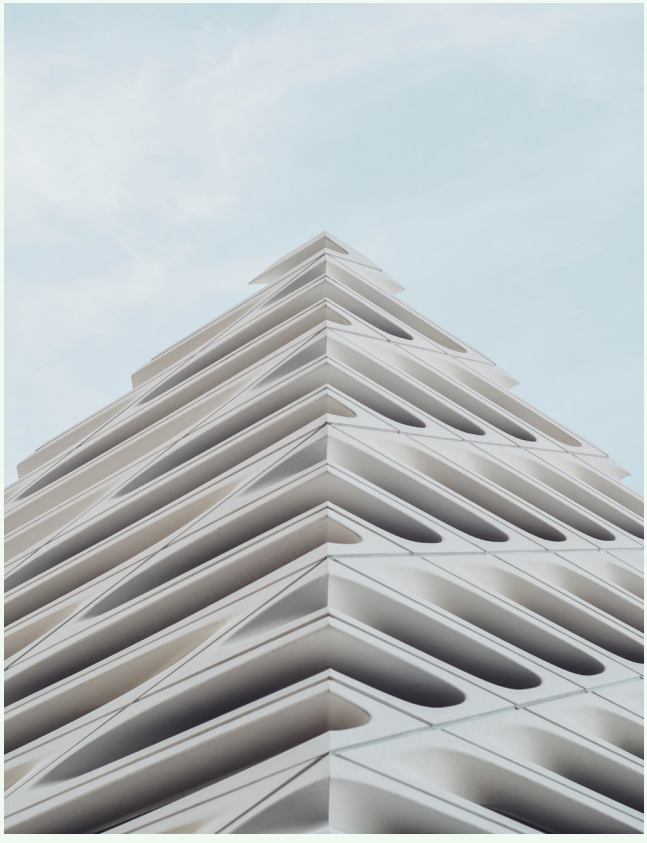
In these cases, a combination of dedicated fibers and microwave systems are often used to shorten the path.

It should be noted that whilst physical distance and delay are critical for these applications, the bandwidth requirements are relatively small.

Lastly, **software-based redundancy** (where content is mirrored across different servers) is very latency-sensitive and servers must be located within 10 milliseconds of each other. This is necessary to achieve adequate physical redundancy on one hand and to meet the stringent latency demands of software-mirroring on the other.

To conclude, if we exclude trading and software mirroring, there are few applications for which the user experience will be adversely affected by the additional latency from remote server location. For the northern European market, most applications can be served efficiently from Stockholm.

www.stockholmdataparks.com
www.teliacarrier.com



[14] http://glinden.blogspot.com/2006/12/slides-from-my-talk-at-stanford.html
[15] https://services.google.com/fh/files/blogs/google_delayexp.pdf
[16] The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency, https://www.pubnub.com/blog/how-fast-is-realtime-human-perception-and-technology/
[17] Latency Thresholds for Usability in Games: A Survey, https://ojs.bibsys.no/index.php/NIK/article/view/9
[18] Kjetil Raaen, Ivar Kjellmo. Measuring Latency in Virtual Reality Systems, https://hal.inria.fr/hal-01758473/document. Towards Low-Latency and Ultra-Reliable VirtualReality, https://arxiv.org/pdf/1801.07587.pdf
[19] https://passive-components.eu/autonomous-vehicles-may-not-rely-on-5g/