

White paper

Latency

The impact of latency
on application performance

Nokia Siemens
Networks



Contents

3 Executive summary

4 1 Latency and the end user

4 1.1 What is latency?

4 1.2 Low latency keeps users happy

5 1.2.1 Web browsing vulnerable to latency

5 1.2.1.1 Web page sizes and traffic going up

5 1.2.1.2 Long response times frustrate users

5 1.2.1.3 Fast response times bring more revenue

6 1.2.2 Online gaming needs low latency

7 2 Latency in mobile broadband

8 2.1 I-HSPA – unique flat network architecture

9 2.2 Multicontroller RNC keeps latency low

9 2.3 LTE – another cut in latency

9 2.4 Fast setup times

10 2.5 Multilayer Optimization keeps transport delay low

10 2.6 Standardization

11 3 Real-life latency measurements

11 3.1 Latency in HSPA networks today

11 3.2 Web page download times

12 3.3 LTE

12 3.4 How to measure network latency

13 4 Conclusion

14 References

14 Glossary



Executive summary

With the introduction of flat rate tariffs and the evolution of technology that is able to provide an acceptable user experience, mobile broadband is experiencing a boom. Mobile broadband also provides a key additional revenue source for communications service providers (CSP) suffering from declining voice revenues. The spread of mobile broadband is also evidenced by the fact that in many developed markets, users get a mobile broadband connection instead of, or in addition to, a fixed broadband connection, meaning that a decent user experience is needed – fast, responsive, always on and available for use.

One of the key ingredients for a good user experience is latency. While the focus in user experience so far has been on maximum bitrates, we show that after a certain level of throughput has been achieved, actually latency - the time from a user sending a piece of data, e.g. requesting a download or a Web page to load, to the time when the user gets a response - is sometimes even more important than the throughput, or bit rate, offered.

Real life examples from the Internet show that an increase in latency can decrease revenues for an Internet service significantly. Some services, such as online gaming, can even be impossible to offer with connections that have a delay.

Modern mobile broadband networks based on HSPA can already offer a decently low latency for most applications. However, low latency is not a given and care must be taken to design the network in a way that provides the lowest latency. When studying networks, we can see that some parts of the network generate most of the latency, with radio access and transport playing key roles in efforts to cut response times on the broadband highway.

In this paper we study the impact of latency on applications, identify the possible bottlenecks for latency in mobile broadband networks and highlight the key developments in mobile broadband technologies that decrease latency.

1. Latency and the end user

1.1 What is latency?

Latency and throughput are the essential factors in network performance and they define the speed of a network. Whereas throughput is the quantity of data that can pass from source to destination in a specific time, round trip time (RTT) latency is the time it takes for a single data transaction to occur, meaning the time it takes for the packet of data to travel to and from the destination, back to the source. Latency is measured in milliseconds (ms).

As most of the end user applications used over the Internet are based on TCP, latency is crucial to the broadband experience. TCP requires the recipient of a packet to acknowledge its receipt. If the sender does not receive a receipt in a certain amount of time (ms), TCP assumes that the connection is congested and slows down the rate at which it sends packets. Short data transfers (also called Web “mice”) suffer more from reduced TCP performance. [1]

High latency causes noticeable delays in, for example, downloading Web pages or when using latency sensitive applications.

1.2 Low latency keeps users happy

Why does latency matter? Many people know the feeling of frustration when waiting for Web pages to open when using a slow connection. Consumers are becoming ever more active in the Internet, forming communities to discuss and compare network performance with the aim of getting more from their favorite online activities.

All services benefit from low latency. System latency is more important than the actual peak data rates for many IP based applications, such as VoIP, music downloads, e-mail synchronization and online gaming

People are becoming more “latency aware” and it is being increasingly evaluated and discussed. Fast response times matter. With low latency, the end user has a better experience of Web browsing, as Web page pages download more quickly. Happy customers mean less churn, while fast response times also mean more revenue.

“Typically, more than 80% of all data bursts in WCDMA/HSPA networks are so small (<100kB) that they are more sensitive to latency than throughput.”
Source: Major global CSP

1.2.1 Web browsing vulnerable to latency

A packet takes time to travel from a server to the client and there is a limited number of packets that can be sent (in a TCP/IP data transfer) before the server stops and awaits acknowledgement for packets already received. This means that excessive latency has a negative effect on transfer speed. Web page download time is affected by the performance of the browser, the speed of the Internet connection, the local network traffic, the load on the remote host and the structure and format of the Web page requested

1.2.1.1 Web page sizes and traffic going up

Within the last five years, the average size of a Web page size has more than tripled, from 100kB

“One must also not ignore the impact of latency when determining the maximum user throughput. For applications, such as email and Internet surfing, which require some form of TCP/IP packet acknowledgements, the theoretical maximum data rate can be limited by the latency of the end-to-end connection.”

Source: Signals Ahead Vol.5, No 4, Signals Research Group, 2009

to more than 300kB, while the use of streaming media on the Web has increased by more than 100% each year. During the same time, the number of objects has nearly doubled, requiring measures to optimize Web performance. [2]

1.2.1.2 Long response times frustrate users

This raises the question of how long are end users willing to wait for a Web page to be downloaded before abandoning the attempt? The end users' TWT, tolerable waiting time, is getting shorter. Studies suggest that feedback during the waiting time, for example, percent-done indicators, encourages users to wait longer. The type of task, such as information retrieval, browsing, online purchasing or downloading also has an effect on a user's level of tolerance.

Users can become frustrated by slow Web sites, which are perceived to have lower credibility and quality – by contrast, faster Websites are perceived to be more interesting and attractive.

1.2.1.3 Fast response times bring more revenue

In a CNET article, Marissa Mayer, Vice President of Search Product and User Experience, Google, said that when the Google Maps home page was put on a diet, shrunk from 100K to about 70K to 80K, traffic was up 10 percent the first week and grew 25 percent more in the following three weeks. [3] “Users really respond to speed”, she says.

Google also found that moving from a 10-result page loading in 0.4 seconds to a 30-result page loading in 0.9 seconds decreased traffic and ad revenues by 20%. It turned out that the cause was not just too many choices, but the fact that a page with 10 results was half a second faster than the page with 30 results.

“Google consistently returns search results across many billions of Web pages and documents in fractions of a second. While this is impressive, is there really a difference between results that come back in 0.05 seconds and results that take 0.25 seconds? Actually, the answer is yes. Any increase in the time it takes Google to return a search result causes the number of search queries to fall. Even very small differences in results speed impact query volume.” [3]

Experiments at Amazon have revealed similar results: every 100 ms increase in load time of Amazon.com decreased sales by one percent. [4] Also, tests at Microsoft on Live Search [5] showed that when search results pages were slowed by one second:

- Queries per user declined by 1.0%, and
- Ad clicks per user declined by 1.5%

After slowing the search results page by two seconds:

- Queries per user declined by 2.5%, and
- Ad clicks per user declined by 4.4%

1.2.2 Online gaming needs low latency

There are millions of Internet users participating in online gaming daily. Broadband Internet has opened up endless possibilities for virtual gaming. Most popular games include Action Games, Adventure, Simulations, Racing and Multiplayer Online Role Playing Games. Generally, network gaming requires very low latency, but the required maximum latency depends on the type of game.

The computer and the game server take time to decode the messages sent to each other and execute an operation. Because of this, gamers experience latency: delay or lag – a slowdown in game play and in some cases the entire game “freezing” for several seconds. RTT delay, jitter and application packet loss are the main performance parameters for defining gaming performance requirements. The first person shooter and MMOG game genres are in general most sensitive to latency.

The user experiences high latency as an inability to react quickly enough, or ultimately losing the game due to being slow. There are several online gaming communities on the Internet which

discuss and compare the latency and network performance of service providers. Obviously, a low latency fast connection will reduce churn. Multiplayer Online Role games have been known to have a subscriber base of around 10 million for one game.

According to a PCMagazine article [6], the total cost to Blizzard of running its game WoW (World of Warcraft) is around USD 200 Million since 2004. With an average monthly fee of USD 15 per month, with 10 million players Blizzard would make USD 150 Million each month with that game. According to these estimates Blizzard could cover their entire costs for the game over a four-year period in less than two months.

Another PC Magazine article [7] refers to analyst estimates of the online game market being at about a fifth of the size of the video console game market. Total PC game revenue is expected to reach USD 19 billion by 2013.

2. Latency in mobile broadband

Previously, we discussed how latency can affect end-user experience and applications. How can we lower the latency to improve the end user experience and provide our customers with the advantage of low latency? Reducing latency in the radio access network makes it possible to provide services which are sensitive to it.

“Conversely, if the latency in a network was reduced by a third then the maximum data rate that the network could deliver would actually triple without making any other changes to the network (eg the introduction of MIMO, higher modulation schemes, etc.)”
Source: Signals Ahead Vol.5, No 4, Signals Research Group, 2009

One of the main drivers in the evolution of 3G, from 3GPP Rel99 to HSPA and further to HSPA+, is to reduce the latency to provide better support to

applications that are sensitive to time delays. HSDPA and HSUPA are the two established standards of HSPA. HSPA was included in 3GPP Releases 5 and 6 for downlink and for uplink with data rates comparable to ADSL-services. The introduction of HSPA to the WCDMA networks is providing a similar user experience to Digital Subscriber Line (DSL) access. HSPA Evolution, also known as HSPA+, is an enhancement of HSPA, which allows 3G CSPs to exploit the full potential of their WCDMA systems.

When the 3GPP worked on defining the requirements for HSPA evolution, the objective was to achieve a round trip time of 50 ms instead of 100 ms. Figure 1 shows today's latency - Nokia Siemens Networks has a measured latency of 41 ms in live networks already with 3GPP Release 6 in RNC based networks. It also shows the expected latency evolution for HSPA and LTE: <25 ms for HSPA+ and <20 ms for LTE.

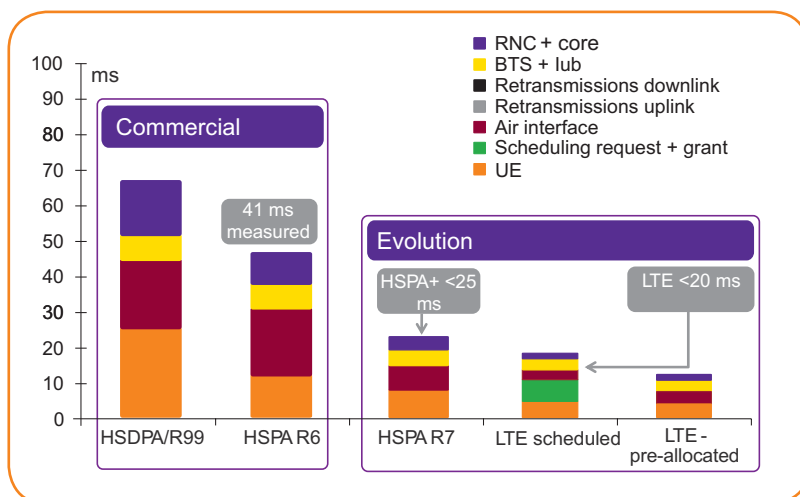


Figure 1: Expected latency evolution for HSPA and LTE

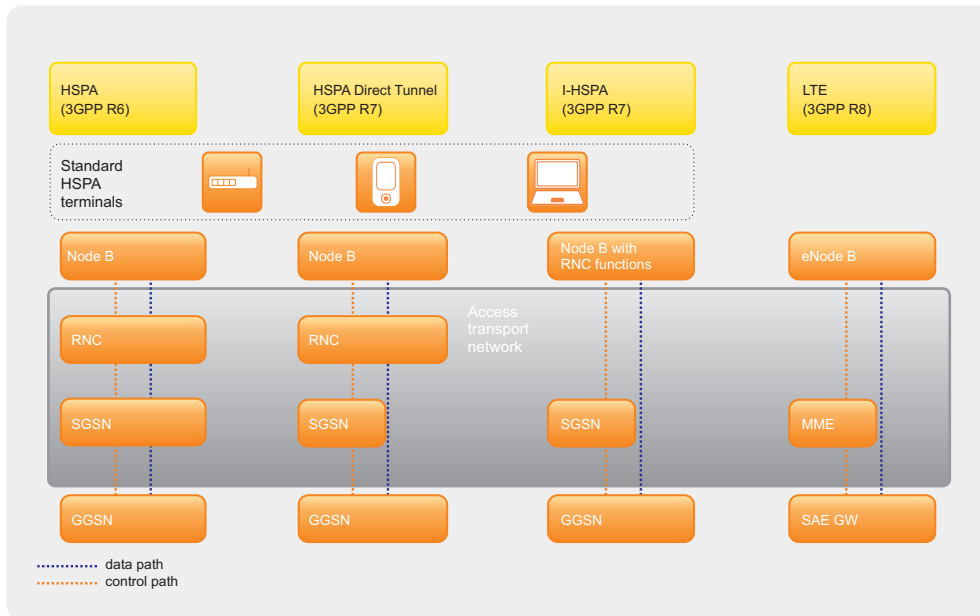


Figure 2: I-HSPA: Less nodes in the data path improves Latency

2.1 I-HSPA – unique flat network architecture

Flat network architecture can further improve end-user experience and CSP efficiency. The first standardized flat architecture in 3GPP networks is in 3GPP R7, also known as I-HSPA. The simplified two-node architecture of I-HSPA (see figure 2) employs a base station that integrates a subset of Radio Network Controller (RNC) functionality and Serving GPRS Support Node (SGSN), supporting the Direct Tunnel feature. This enables data traffic to by-pass the RNC and SGSN and network elements to connect the Radio Access Network (RAN) to the core packet network directly.

With fewer network elements, Nokia Siemens Networks I-HSPA can reduce the RTT further. The I-HSPA flat architecture brings significant improvements, particularly for the RTT measured in IDLE state, because there is no lub related transport set-up delay. Instead, the transport connection is on all the time and the IP packets are sent immediately.

“It’s been a great relationship. What we found with Nokia Siemens Networks I-HSPA solution was an opportunity to deploy a network that was extremely competitive, very cost-effective and very reliable. It flattened the network architecture and eliminated the latency issue. What that means to us as a service provider is we now have the ability to go layer on VoIP solutions and other traditional ISP services on the network.”
Ed Evans, CEO & President, Stelera Wireless

2.2 Multicontroller RNC keeps latency low

The compact architecture in Nokia Siemens Networks Multicontroller RNC, achieved through co-location of UP and CP, will reduce need for node internal traffic inside the RNC and hence reduce latency to a negligible level.

The new platform, which will also be applied to future adapters, is built on the common processing environment for all the functionalities, allowing more flexibility in the way resources are used.

The Multicontroller RNC will cut RNC latency still further.

2.3 LTE – another cut in latency

Evolution to LTE in 3GPP Release 8 brings a further cut in latency, to <20 ms. Measurement results can be found in Chapter 4: Real-life Latency measurements. The LTE physical layer uses a very short frame size of 1 ms, designed to enable a very low latency.

2.4 Fast setup times

We have been discussing latency during a connection, yet another element of latency is connection setup time. Faster connection times and low latency lead to higher effective data rates. Fast setup times - lead to 50-100% higher effective data rates

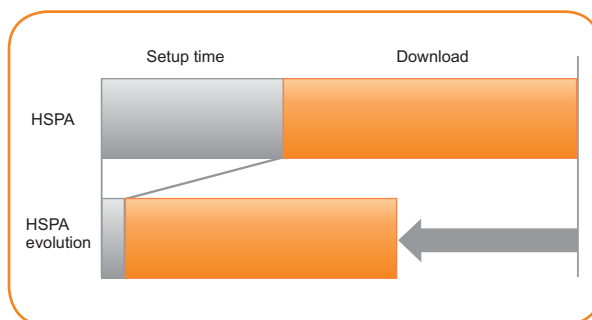


Figure 3: Fast set up time improves end user performance

The channel setup time is typically 0.5-1.0 s in an HSPA network. HSPA evolution enhances the end user performance by reducing the setup time to below 0.1 s. HS-FACH/HS-RACH, defined in 3GPP R7, gives access to high data rates with less delay, achieving better throughput for short connections. Fast setup time and low latency also means that applications originally designed for the wired Internet can now also give a good end-user experience on mobile devices in a HSPA network.

2.5 Multilayer Optimization keeps transport delay low

With the advent of every new radio access technology, such as HSPA+ or LTE, a good deal of effort is expended to reduce the system inherent delay. Hence, as the system inherent delay decreases, the transport delay becomes more relevant to the service quality as perceived by the end user. The following looks at how some key transport developments can decrease latency.

Looking at today's situation, the backhaul network is particularly bandwidth-constrained. Often these networks have been inherited from 2G and have been given only minor upgrades during the introduction of 3G. As a consequence, links have to be oversubscribed, meaning that low priority traffic is being queued at transport nodes. In 2G and 3G it is a system-inherent requirement to treat signaling and voice traffic with the highest priority. As a consequence, HSDPA is the traffic type that is oversubscribed most aggressively and consequently queued.

Our recommendation is to modernize the transport network. Carrier Ethernet is our proposition for the backhaul network, with IP/MPLS (Multiprotocol Label Switching) for the backbone network. Carrier Ethernet widens access bottlenecks and introduces traffic multiplexing close to base stations. It also introduces Gigabit Ethernet and 10 Gigabit Ethernet rings in the aggregation network. Carrier Ethernet smoothly plugs into our IP/MPLS based backbone designs, allowing CSPs to prepare for flat architectures (I-HSPA, LTE) today.

One of the key values of our Carrier Ethernet solution is Quality of Service. Connections are traffic-engineered end-to-end, ensuring minimum queuing (or optionally no queuing at all) along the entire transport chain. This allows Carrier Ethernet to be applied in environments that were previously

thought to require a dedicated fiber connection. Multilayer Optimization is a Nokia Siemens Networks' concept for processing traffic on the lowest layer possible. This can often be the optical layer - traffic is switched directly at the photonic layer ("agile photonic networking") instead of being put through switches and routers. Dense Wavelength Division Multiplexing (DWDM) is the technology used for this. Both our Carrier Ethernet as well as our IP/MPLS solutions make extensive use of DWDM, which is tightly integrated to both.

Nokia Siemens Networks' mobile network elements are able to indicate traffic priorities to the transport layer. This ensures a minimum amount of queuing for the most delay-sensitive traffic, leading to the lowest possible latency. This priority indication can be coarse, for example voice, PS data, HSDPA, or fine-granular, for example per user, who can be classified gold / silver / bronze.

Our integrated approach, bringing radio and transport together, ensures that the priorities indicated by the mobile network elements are consistently applied along the entire transport chain, particularly in the bandwidth-constrained backhaul network.

2.6 Standardization

I-HSPA, High speed FACH and High speed RACH are some of Nokia Siemens Networks' work items. HS-FACH and HS-RACH are HSPA Evolution features giving access to high data rates without any setup latencies.

3. Real-life latency measurements

This chapter presents some Nokia Siemens Networks real-life latency measurement results.

3.1 Latency in HSPA networks today

In a latency measurement done with major European CSPs in commercial HSPA networks, Nokia Siemens Networks has 40% lower latency than other vendors.

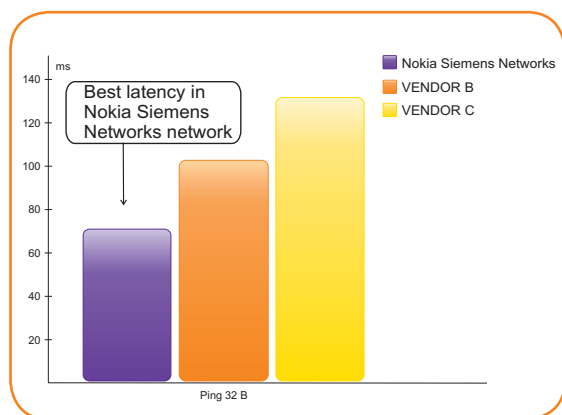


Figure 4: Measured HSPA latency. All networks had 7.2 Mbps HSDPA and HSUPA enabled.

Also, an average RTT down to 41 ms has been measured in Nokia Siemens Networks' live networks with 3GPP Rel 6 (see Figure 1 in Chapter 3).

3.2 Web page download times

Web page download tests were run with the same CSPs, comparing times to the same vendors. The low latency affects Web page download time directly.

Medium to large size Web pages:

- Nokia Siemens Networks: ~ over twice as fast as Vendor B
- Nokia Siemens Networks: ~ over 3 times faster than Vendor C

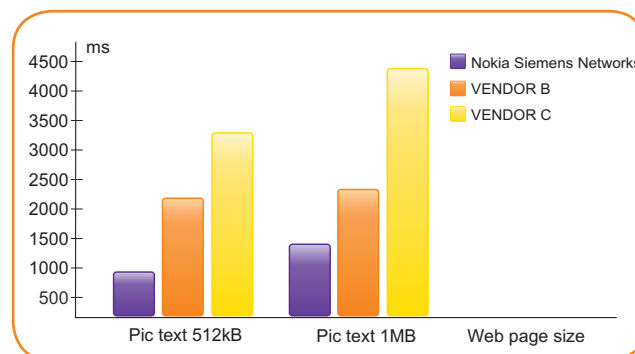
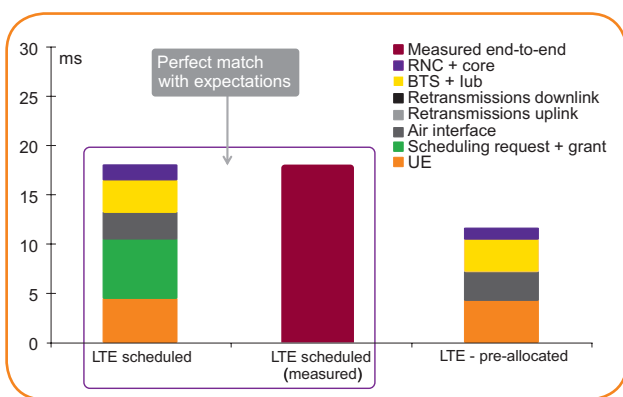


Figure 5: Average HTTP download times.

Latency and Web page download time were measured in commercial HSPA networks. All networks had 7.2 Mbps HSDPA and HSUPA enabled.



Scheduled: UE requests for resources, which BTS allocates. After that UE is able to transmit.

Pre-allocated: BTS has pre-allocated resources to UE, UE can transmit immediately.

Figure 6: LTE bringing latency further down- Nokia Siemens Networks LTE latency measurements

3.3 LTE

The measurements of Nokia Siemens Networks' LTE end-to-end latency, from a commercial trial with a major global CSP, show a perfect match with expectations. Measurements show a stable ping below 20 ms.

3.4 How to measure network latency

Ping tests measure latency by determining the time it takes a given network packet to travel from source to destination and back, for example, speed of information traveling from a computer to a game server. A ping estimates a round trip time using interval timing responses. An average response time provides a good measure of the end-to-end round-trip time through the network between terminal and server in the data network. For repeatable and comparable RTT results, an RTT measurement protocol is recommended and is used in all Nokia Siemens Networks' measurements.

4. Conclusion

Reducing latency in the radio access network makes it possible to provide delay sensitive services, offering CSPs new revenue possibilities. Through optimizing the network end-to-end and reducing network element latency, it is possible to get good results, as is shown by some real-life measurement results in the HSPA networks of today. Optimizing transport networks is also a key to decreasing latency.

Evolution to LTE is decreasing latency still further, with measurement results showing a perfect match with expectations.


We can say that latency is relatively more important to the consumer mobile broadband experience than data rates. With low latency, a CSP can increase end-user satisfaction and thus reduce churn.

Glossary

3GPP	3rd Generation Partnership Project
ADSL	Asymmetric Digital Subscriber Line
BTS	Base Transceiver Station
CP	Control Plane
CSP	Communications Service Provider
DSL	Digital Subscriber Line
DWDM	Dense Wavelength Division Multiplexing
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
HSDPA	High Speed Downlink Packet Access
HS-FACH	High Speed Forward Access Channel
HSPA	High Speed Packet Access
HS-RACH	High Speed Random Access Channel
HSUPA	High Speed Uplink Packet Access
I-HSPA	Internet- High Speed Packet Access
IP	Internet Protocol
LTE	Long Term Evolution
MPLS	Multiprotocol Label Switching
MIMO	Multiple-input multiple-output
MMOG	Massively multiplayer online game
PS	Packet switched
QoS	Quality of Service
RAN	Radio access network
RNC	Radio network controller
RTT	Round Trip Time
SGSN	Serving GPRS Support Node
TCP	Transmission Control Protocol
TWT	Tolerable waiting time
UP	User plane
VoIP	Voice over Internet Protocol
WCDMA	Wideband Code Division Multiple Access

References

- [1] Lian Guo& Ibrahim Matta: "The War between Mice and Elephants" Boston University Technical Report, 2001
- [2] Betcher, Bill and Flinn, David of Gomez Inc: "Top 1000 Home Page Data from Gomez, Inc", 2008
- [3] CNET News.com/ZDNET.com: "Google's Marissa Mayer: Speed wins" <http://blogs.zdnet.com/BTL/?p=3925> by Dan Farber, November 9, 2006. Google's Marissa Mayer: Speed wins. http://www.google.co.uk/enterprise/end_user_experience.html
- [4] Kohavi Ron, Henne Randal M, Sommerfield Dan of Microsoft: "Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO", 2007.
- [5] Kohavi Ron, Longbotham Roger, Sommerfield Dan of Microsoft: "Controlled Experiments on the Web: Survey and Practical Guide", 2009
- [6] PC Magazine: 09.16.08, <http://www.pcmag.com/article2/0,2817,2330507,00.asp>
- [7] PC Magazine 02.02.09, <http://www.pcmag.com/article2/0,2817,2339966,00.asp>



Nokia Siemens Networks Corporation
P.O. Box.1
FI-020022 NOKIA SIEMENS NETWORKS
Finland

Visiting address
Karaportti 3, ESPOO, Finland

Switchboard +358 71 400 4000 (Finland)
Switchboard +49 89 5159 01 (Germany)

The contents of this document are copyright © 2009 Nokia Siemens Networks. All rights reserved.

A license is hereby granted to download and print a copy of this document for personal use only. No other license to any other intellectual property rights is granted herein. Unless expressly permitted herein, reproduction, transfer, distribution or storage of part or all of the contents in any form without the prior written permission of Nokia Siemens Networks is prohibited.

The content of this document is provided "AS IS", without warranties of any kind with regards its accuracy or reliability, and specifically excluding all implied warranties, for example of merchantability, fitness for purpose, title and non-infringement. In no event shall Nokia Siemens Networks be liable for any special, indirect or consequential damages, or any damages whatsoever resulting from loss of use, data or profits, arising out of or in connection with the use of the document. Nokia Siemens Networks reserves the right to revise the document or withdraw it at any time without prior notice.

Nokia Siemens Networks and the Wave-logo are registered trademarks of Nokia Siemens Networks. Nokia Siemens Networks product names are either trademarks or registered trademarks of Nokia Siemens Networks. Other product and company names mentioned herein may be trademarks or trade names of their respective owners.