HELSINKI UNIVERSITY OF TECHNOLOGY
Networking Laboratory

Mobile Operator Measurements: Case COIN

Antero Kivi

25.4.2007

Agenda

- Introduction
- Measurement process
- Processing of data
- Other issues
- Exemplary results

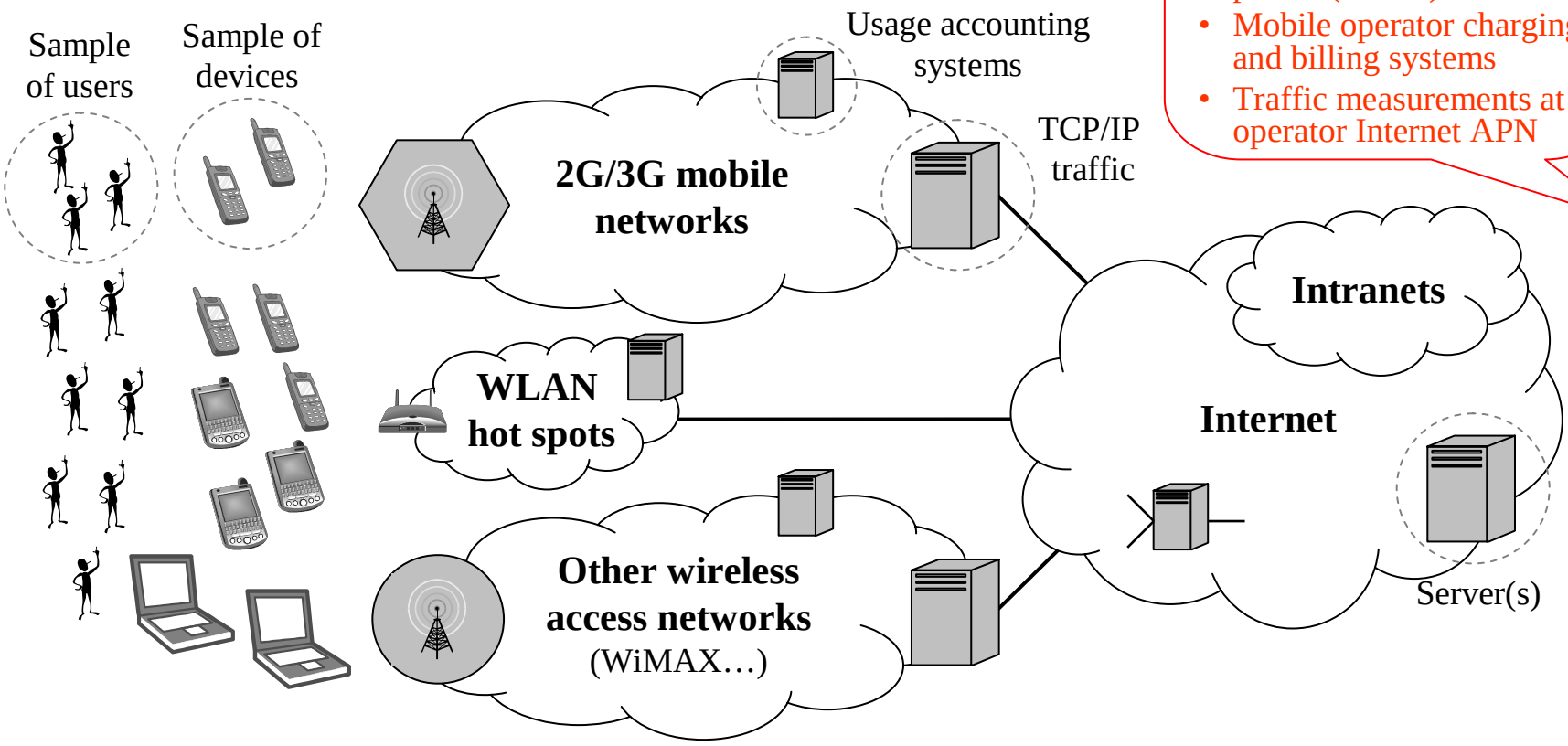
Introduction

- COIN - Dynamics of COmpetition and INnovation in the converging Internet and mobile networks
  - *COIN aims to improve the techno-economic understanding of the dynamics of the national mobile and wireless services market. The key stakeholders of the Finnish mobile market are involved in the project, and the results will contribute to both regulatory and strategic planning.*
- COIN research partners
  - Tekes
  - Nokia
  - TeliaSonera
  - Elisa
  - DNA Finland
  - Digita
  - Ministry of Transport and Communications
- COIN research activities
  - Service usage measurements
  - Handset bundling study
  - Market effects of emerging radio technologies
  - Disruptive applications and services
  - Ecosystem structure study
  - Mobile Operator Business game

Sources of data on mobile service usage

OUR RESEARCH

- Surveys on handset panel participants
- Handset monitoring panels (SP360)
- Mobile operator charging and billing systems
- Traffic measurements at operator Internet APN

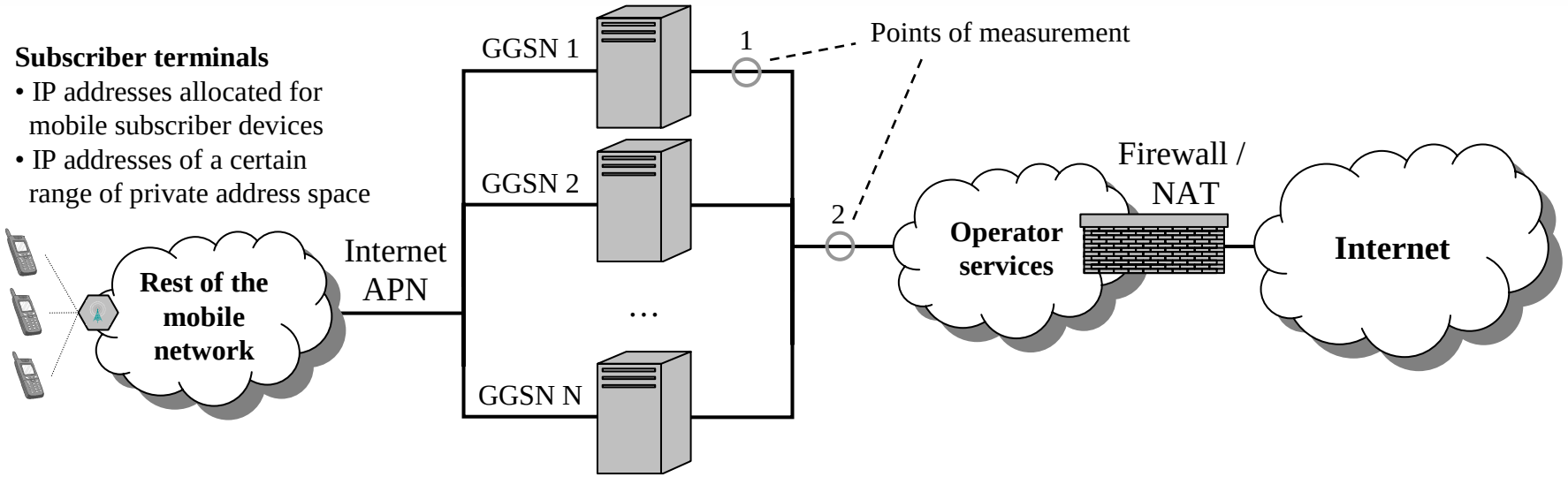


Source: Kivi, 2007

Measurement objectives and scope

- Why traffic measurements?
  - To understand mobile data service usage and end-user behavior
  - To see mobile data traffic profile beyond data volumes
  - To relate observations to data from other measurements
- What is measured?
  - All packet data traffic of Finnish mobile terminals to/from the Internet during a specified time period targeted
- Measurements at three major Finnish mobile operators' networks
  - TeliaSonera, Elisa, DNA (no Saunalahti, TeleFinland)
  - 80-90% of all Finnish mobile subscribers/terminals
  - → represents the whole market
- One-week period from all operators
  - September – October, one week in 2005, two weeks in 2006
  - Measurements not completely simultaneous
- Longitudinal data
  - Measurements conducted in 2005, 2006, 2007…?

Measurement setup



- Centralized point for measurements available at the GSM/UMTS network
  - Internet APN, WAP APN, MMS APN, corporate APN(s)…
  - Internet APN about 90% of total packet data traffic volume in mobile network
  - Includes traffic from postpaid and prepaid subscribers, business and consumer subscribers
- No influence from roaming, as home GGSN roaming is used
  - All roaming traffic by operators' subscribers routed via home network GGSN → all packet data roaming traffic by operators' subscribers included, no foreign roamers' traffic included
- Mobile terminals separated from servers with terminal IP address ranges

Measurement and traces

- Internet APN as the point of measurement
  - Either one, or several GGSNs
  - After the GGSN, before operator servers or NAT/firewall
- TCP/IP headers captured with tcpdump
  - Three passive, single-point measurements
- Resulting in packet level traces
  - TCP/IP headers (SYN,FIN,RST) and all UDP/IP headers (2005)
    - tcpdump -w trace_file.pcap 'tcp[tcpflags] & (tcp-syn|tcp-fin|tcp-rst) != 0 or udp '
  - All above IP headers (2006)
    - tcpdump -w trace_file.pcap ip
  - No application protocol headers captured
    - Could contain sensitive information (email addresses, phone numbers…)
    - Would increase resource needs (disk space, processing time…)
- Large trace files
  - In total >200GB of packet level traces
  - → need for compression/aggregation

Sanitation of trace data

- Remaining application layer data removed
  - Tcpdump does not see the headers, but captures by default the 68 first bytes of the packet → little application layer data might still included in the traces
  - Such data removed with tcpdpriv by the operators (http://ita.ee.lbl.gov/html/contrib/tcpdpriv.0.txt)
- No need to anonymize IP addresses
  - Dynamic addressing used for subscriber terminals
  - Server addresses specifically studied

Processing the traces

- Terminals OSs identified from packet traces
  - p0f passive OS fingerprinting tool (http://lcamtuf.coredump.cx/p0f.shtml)
  - → separate file with an OS entry for each terminal IP and time period
- Packet traces aggregated to flow level
  - CoralReef software suite for collecting and analyzing passive traffic traces (http://www.caida.org/tools/measurement/coralreef/)
  - Separate script for 2005 TCP data
  - → flow level traces
- Further aggregation with perl script(s)
  - OS information combined with flow level traces
  - For all traffic
    - Bytes, flows, packets per OS, day+hour, proto, server port
  - For web traffic (server port 80, 8080, 8000, 8888, 443)
    - Bytes, flows, packets per OS, day+hour, server IP
  - → output files per script
- Output files input to SPSS for further aggregation and plotting

- Domain names for all web server IP addresses resolved
  - All web server IPs extracted
  - Domain names for the IPs resolved with another script

Identification of terminal operating systems

- Terminal operating system identified using *TCP fingerprinting*
  - Differences in implementation of TCP/IP stack in different OSs (e.g. default values of TCP window size, initial TTL, don't fragment bit…)
  - → distinct TCP "fingerprints"
  - Traffic traces are compared to the fingerprints of previously identified OSs
- Operating system identification process includes some possible bias
  - Only mobile terminal OSs identified (i.e. with IP addresses in a specified range)
  - OS identification is based on uplink TCP traffic only (57% of flows, 23% of bytes)
    - OS of uplink TCP flows identified → that OS resides at a certain IP address at a certain time frame
    - Downlink TCP flows, and all UDP flows accounted for different OSs based on this information
  - Fingerprint database is not complete
    - Common PC and smart phone OSs can be identified with reasonable accuracy
    - Fingerprints of 15-20 Symbian handsets submitted to the database by TKK to improve handset identification

Sensitivity issues

- Legislative issues
  - No application layer data
    - No phone numbers, email addresses…
  - Dynamic IP addresses for subscriber terminals
    - Subscriber can't be identified from the address
  - → No identification information in trace data, data deliverable to TKK
  - See lecture "Traffic measurements and Finnish legislation"

- Business sensitivity
  - Not to reveal anything to competitors, nothing operator-specific
  - Sum of 3 operators → can't single out any one operator from the results

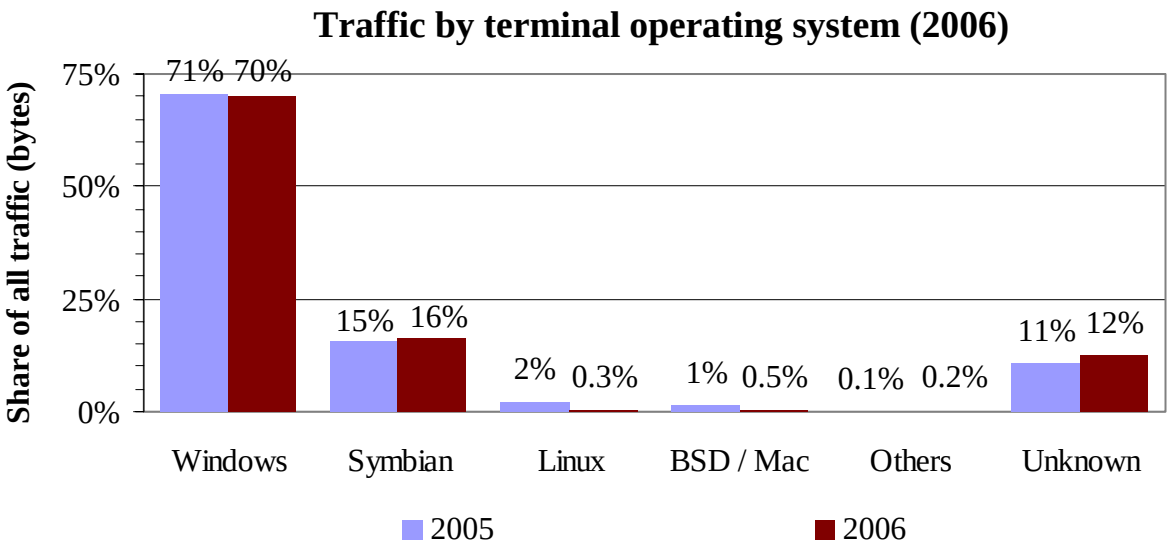
Practical learnings

- Biggest job is in selling the idea to the operator
  - Not just to operator contact person, but also inside operator organization to people actually implementing the measurement (and their bosses)
  - Cost/benefit? What's the output?
  - Are there legal obstacles?
  - Can TKK be trusted with our data? NDAs?
  - Has been an issue in the Funet measurements as well
- The actual measurement is technically straightforward
- Resource requirements big, and increasing
  - Not enough disk space in trace collection
  - Not enough memory in analysis
  - Traffic volume increasing all the time (4x between 2005 and 2006)
    → even more resources needed next year

Exemplary results

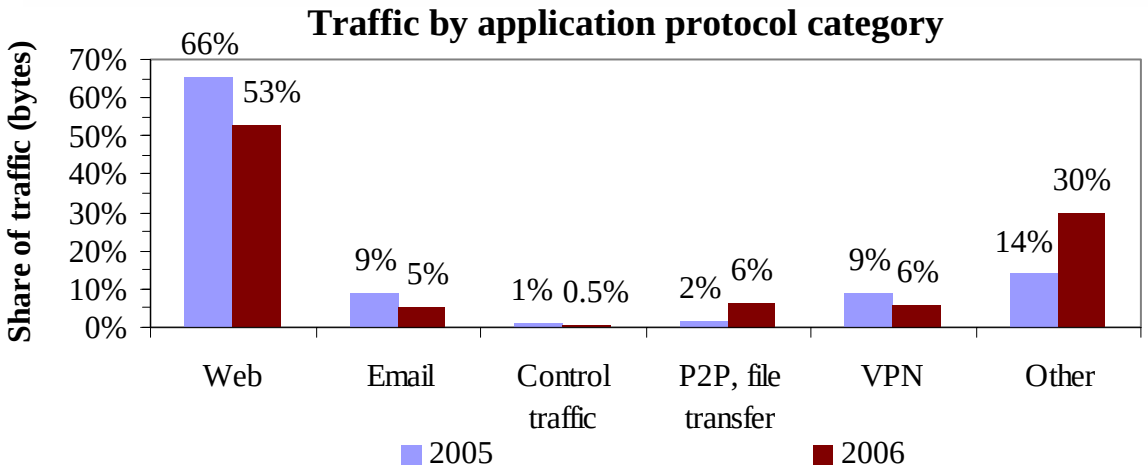
- Traffic by mobile terminal operating system
- Traffic by application protocol category
- Traffic by application protocol category and OS
- Traffic by day and hour
- Most popular web sites by category

Traffic by mobile terminal operating system



- Windows originates 70% of traffic in mobile network
  - Data cards, GPRS modems, handsets as modems via Bluetooth/cable
  - A few PCs create more traffic than many mobiles → OS identification necessary to uncover handset traffic
  - Windows Mobile, Windows CE, and Pocket PC traffic in "Others" category
- At least 16% of traffic actually made with Symbian handsets
  - 32% of traffic with Symbian device as the GSM/UMTS network terminal (CDR data) → 4–16% of this traffic from modem usage
- Unknown 12% of traffic problematic
  - All other handsets, possibly additional laptop and Symbian traffic
  - Telematics, machine-to-machine (M2M) comm., alarm terminals, remote cameras…?
  - Do intelligent modems / GPRS modules, VPN, or access network elements (SGSN/GGSN, firewall) alter the TCP fingerprint?

Traffic by application protocol category

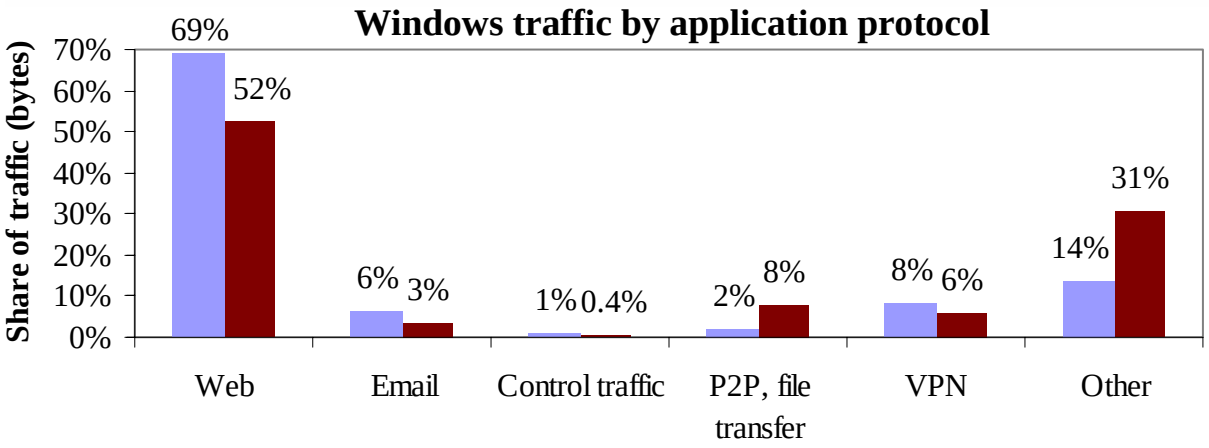


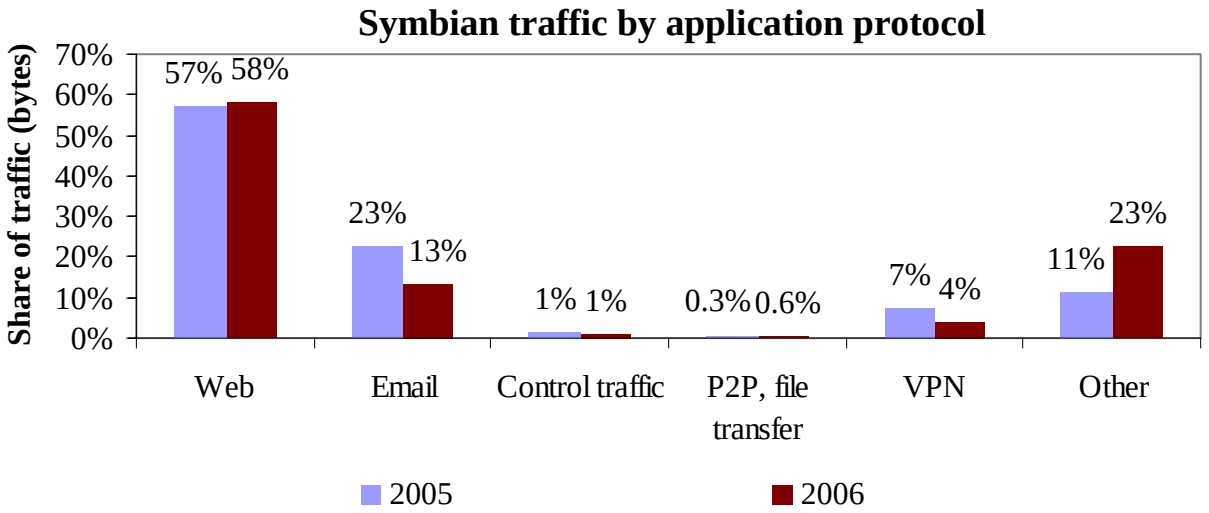
Application protocol category	Transport protocol ports included*	
Web	TCP	80, 443, 8080
Email	TCP	110, 143, 25, 993, 995
Control traffic	UDP	53, 5060, 123
P2P, file transfer	TCP	4662, 7777, 6881, 1412, 20, 9999, 6346, 411, 6882, 412
VPN	TCP	10000, 500
	UDP	2746, 10000, 4500, 500, 1194
Others	TCP	119, 7171, 11469, 554, 1863, 12001, 1352, 3000, 1935, 32459, 22, 1750
	UDP	370, 32555, 5004

* TCP/UDP ports with at least 0,5% of total bytes in category

- Application protocols identified with server-side TCP/UDP port numbers
  - Nearly all 65000 TCP and UDP ports observed
  - Port number based identification not full proof
- Port numbers grouped into 6 application protocol categories to simplify
- Web traffic dominates with >50% traffic share
- ”Other” category with most growth
  - P2P and streaming becoming more mainstream also in mobile network?
  - Also client ports, self-initiated Windows traffic, client ports, malware
- P2P traffic increasing
  - More P2P likely in Web and ”Other”
  - Still much less than in fixed Internet

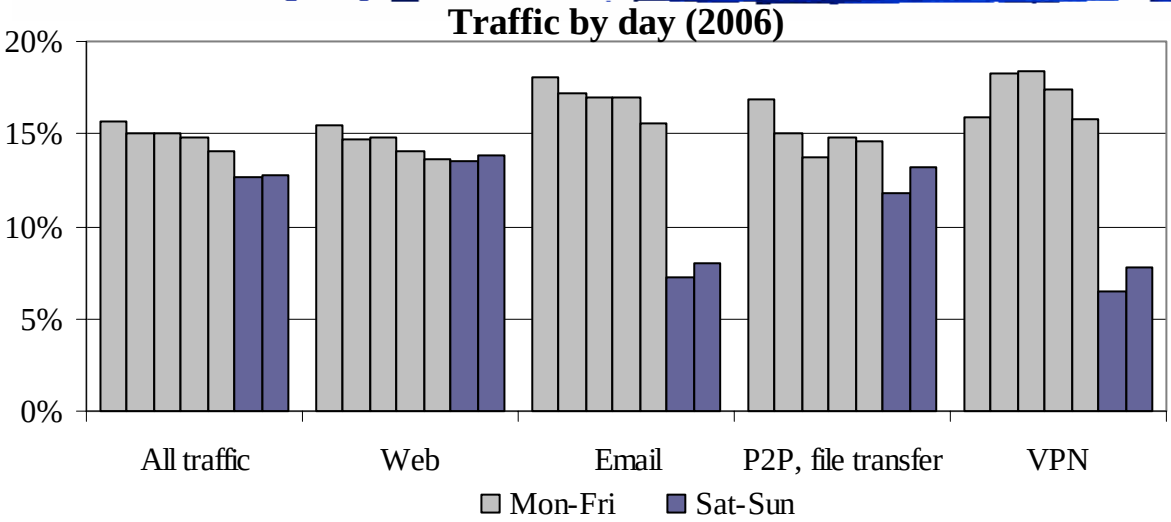
Traffic by application protocol category and OS

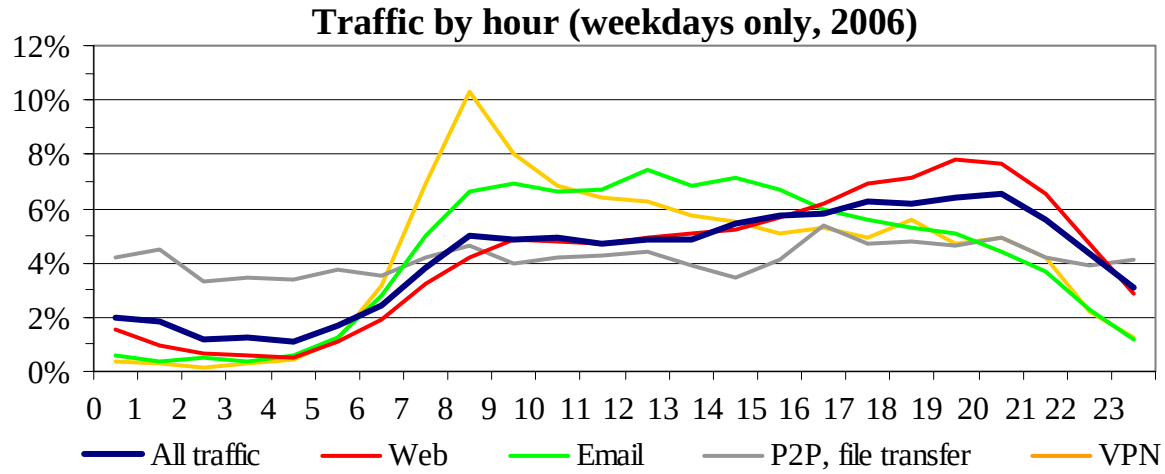




- Windows profile as the profile for all traffic
  - Imposes itself with 70% share of traffic
- Symbian profile differs from Windows in some ways
  - Email share 3x higher
  - P2P share only 1/10

Traffic by day and hour





- Monday most active day, peak hour at 8–9PM
  - 16% of weekly traffic on Mondays
  - 6.5% of daily traffic at 8–9PM
- Business-oriented email and VPN less used on weekends
- Web is free-time oriented
  - Almost equal usage during weekdays and weekend
  - Most usage in the evening on weekdays

Most popular web sites by category

Rank*	Site category	Major domains included**	Share of web traffic volume	Share of web site visits
1.	Mobile operator site	-	11.8%	9.9%
2.	Information	mtv3.fi, sanomawsoy.fi, iltalehti.fi, yle.fi, sanoma.fi, helsinginsanomat.fi	8.8%	11.8%
3.	Entertainment	irc-galleria.net, youtube.com, telkku.com, veikkaus.fi	5.9%	5.7%
4.	Web search	google.com, yahoo.com	2.5%	3.0%
5.	Adult content	seksitreffit.fi, sihteeriopisto.net	1.8%	1.7%
6.	Messaging	luukku.com, hotmail.com, msn.com, passport.net, gmail.com	1.4%	2.9%
7.	Banking	op.fi, sampo.fi, nordea.fi, eQonline.fi, huoneistokeskus.fi, alandsbanken.fi, aktia.fi, osuuspankki.fi	1.4%	1.2%
8.	Advertising	doubleclick.net, advertising.com, adtech.de, theonlinetrader.com, tradedoubler.com	1.2%	3.4%
	Hosting / corporate site	akamai.net, statistik-gallup.net, basefarm.net	23.3%	22.2%
	Other sites	-	12.5%	10.2%
	Unknown	-	19.2%	18.8%
	Private	-	10.2%	9.2%

- Server IP addresses of all TCP flows with ports 80, 8080, 8000, 8888, and 443 included
  - Might include e.g. P2P traffic as well
  - → 225000 web server IP addresses
  - → 41000 domain names
  - Domain names grouped into 12 categories, despite many potential sources of error
- Mobile operator sites (12%), information (9%) and entertainment (6%) significant categories
- Lots of traffic to “infrastructure” hosts not intentionally connected by users
  - Advertising (banners, pop up windows)
  - Load sharing (e.g. akamai.net)
  - Web site analytics (e.g. statistik-gallup.net)

* Ranked by the category's total share of traffic volume
** Sites with at least 10% of the total bytes or flows of the category
*** Share of TCP flows to/from the domain: # of web site visits <= # of flows <= files downloaded from site