Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Computer Networks 51 (2007) 1205–1231
Computer Networks
www.elsevier.com/locate/comnet

Diagnosis of capacity bottlenecks via passive monitoring in 3G networks: An empirical analysis

Fabio Ricciato [a], Francesco Vacirca [a,b,*], Philipp Svoboda [a,c]

[a] ftw. Forschungszentrum Telekommunikation Wien, Donau-City-Strasse 1, A-1220 Vienna, Austria
[b] Infocom Department, University of Roma, "La Sapienza", Via Eudossiana 18, 00184 Roma, Italy
[c] INTHFT Department, Vienna University of Technology, Karlsplatz 13, A-1040 Wien, Austria

Received 27 September 2005; received in revised form 23 June 2006; accepted 12 July 2006
Available online 23 August 2006

Responsible Editor: X.S. Shen

Abstract

In this work we address the problem of inferring the presence of a capacity bottleneck from passive measurements in a 3G mobile network. The study is based on one month of packet traces collected in the UMTS core network of mobilkom austria AG & Co KG, the leading mobile telecommunications provider in Austria, EU. During the measurement period a bottleneck link in the UMTS core network was revealed and removed, therefore the traces enable the accurate analysis and comparison of the traffic behavior in the two network conditions: with and without a capacity bottleneck.

Two approaches to bottleneck detection are investigated. The first one is based on the signal analysis of the marginal rate distribution of the traffic aggregate along one day cycle. Since TCP-controlled traffic dominates the overall traffic mix, the presence of a bottleneck strains the aggregate rate distribution and compresses it against the capacity limit during the peak hour. The second approach is based on the analysis of several TCP performance parameters, e.g. estimated frequency of retransmissions. Such statistics are unstable due to the presence of few top users, but this effect can be counteracted with simple filtering methods. Both approaches are validated via simulations.

Our results show that both approaches can be used to provide early warning about future occurrences of capacity bottlenecks, and can complement other existing monitoring tools in the operation of a production network.

Keywords: Traffic analysis; Passive monitoring; Bottleneck detection; 3G; UMTS

1. Introduction

Public wide-area wireless networks are now migrating towards third-generation (3G) systems designed to support packet-switched data services. Europe has adopted the Universal Mobile Telecommunication System (UMTS), developed by 3GPP as an evolution of GSM. A 3G mobile network

* Corresponding author. Address: Infocom Department, University of Roma, "La Sapienza", Via Eudossiana 18, 00184 Roma, Italy. Tel.: +39 06 47852306; fax: +39 06 4744481.
E-mail address: vacirca@infocom.uniroma1.it (F. Vacirca).

1389-1286/$ - see front matter
doi:10.1016/j.comnet.2006.07.011

includes two main sections: a packet-switched Core Network (CN), which is based on IP, and a Radio Access Network (RAN). Along with the UMTS RAN (UTRAN) based on W-CDMA, several operators maintain a parallel GPRS/EDGE RAN evolved from the legacy GSM radio access. A general overview of the GPRS/UMTS network structure can be found in [17]. Evolved cell-phones and PDAs already support a broad range of data services, including traditional Internet applications like e-mail, Web, etc. while 3G interface cards for laptop are available since 2004, often coupled with flat-rate subscriptions. The growing popularity of 3G mobile terminals and services has extended the coverage of Internet access to the geographic area, and 3G networks are becoming key components of the global Internet in Europe [32].

At present the 3G environment is under evolution: the subscriber population and the traffic volume are still in a growing phase (for some details on the network under study see [15]); the relative distribution of terminal types (e.g. laptops vs. handsets) and their capabilities is changing quickly; the portfolio of services that are offered by the operators evolves rapidly. All these aspects build a potential for changes in the traffic pattern that can occur at the macroscopic scale (network-wide) and in a relatively short time frame. In such a scenario, the ability to accurately and extensively monitor the current network state and to early detect global performance drifts and localized anomalies is a fundamental pillar of the network operation and engineering processes. Among other problems it is important to early detect emerging bottlenecks, i.e. network elements that are under-provisioned and therefore cause performance degradation to some traffic aggregate as its volume increases. A bottleneck in the RAN can be a geographical area that is frequently overloaded because the provisioned radio capacity does not match the peak-hour traffic demand. A bottleneck in the CN is often a link or other network element with insufficient capacity to carry peak-hour traffic. A capacity bottleneck always impacts a certain traffic aggregate rather than isolated flows – e.g. all the traffic directed to a certain radio area, or routed over a certain network element.

Monitoring a wide-area network involves considerable costs: the number of links to be monitored is large, and they are spread over a wide geographical area. For some monitoring applications it is necessary to access configuration parameters (e.g. provisioned link bandwidth), logs and built-in counters from several network elements, and considerable costs, complexity and complications are found in practice where it comes to extraction, gathering and correlation of such heterogeneous data – in the format and semantics – from different elements, with different software and from different vendors. In summary, installing and maintaining a monitoring infrastructure with the same capillarity of the production network might be very expensive. On the other hand, the structure of a 3G mobile network is highly hierarchical and centralized: the whole traffic is concentrated in the CN and there are only a few gateway nodes (named GGSNs) that connect it to external networks like the Internet, therefore the whole traffic can be captured at a few monitored links.

The ultimate goal of our research is to develop methods to infer the presence of performance bottlenecks in the network from the analysis of the traffic captured at few centralized monitoring points, without direct access to all network links. A possible approach is to observe the TCP behavior: since TCP is closed-loop controlled its dynamics and performances are dependent on the state of the whole end-to-end flow path. In principle it should be possible to infer the presence of performance bottlenecks by looking at the evolution of the TCP aggregate and/or to individual connections at some point along the path. We devised two possible strategies to diagnose the presence of a bottleneck on some network elements from the analysis of the traffic aggregate routed through it but observed at a different point:

- From the analysis of the aggregate traffic rate.
- From the analysis of TCP performance indicators like frequency of retransmissions and/or round trip time statistics.

In this paper we explore both approaches. This study is based on several weeks of packet traces collected by passive monitoring the core network links of a large operational 3G network (mobilkom austria AG & Co KG, the leading mobile operator in Austria, EU). In the middle of the measurement period a bottleneck link within the UMTS Core Network was detected and removed by a capacity upgrade. Since the traces are complete – i.e. all packet headers were captured and timestamped – it is possible to analyze and compare several aspects of the traffic dynamics in the two network