Next Generation Network Overload Control and Test Bed

P K Beaumont† and M Rio‡

† University College London & British Sky Broadcasting, ‡ University College London

Abstract: This paper provides an overview for a Communication Provider (CP) Next Generation Network (NGN) platform of what is Overload, what causes Overload and what Overload Control Protection (OCP) mechanisms are available to mitigate the impact of these. It also provide an overview of a Test Bed comprising a Test Harness (TH) in development to characterise dynamic behaviour of System Under Test (SUT), specifically an IP Multimedia Subsystem (IMS) with PSTN Emulation Subsystem (PES) providing Publically Available Telephony Service (PATS), on efficacy of current versus novel overload control algorithm(s) on selected interface(s), ie Session Initiation Protocol (SIP).

1. Introduction.

Modern telecommunications infrastructure is based on computing technology. The resources associated with these are finite. Under certain conditions it is possible for these to be exhausted resulting in overload with either loss and/or degradation of service/platform availability, performance and/or integrity. Within component functions/interfaces mechanisms have been implemented that provide overload control. However, there are limitations associated with scope and efficacy of these. This could be an issue, eg for PATS where for a period during a televote a user may be unable to make emergency call because of excessive calling associated with televote denying them access to dial tone.

2. NGN.

NGN [1] platforms are increasingly being deployed by CPs to provide services based on IMS architectural framework defined by 3rd Generation Partnership Programme (3GPP) and European Telecommunication Standards Institute (ETSI) Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN) for mobile and fixed operators respectively.

IMS is based on relatively new protocol interfaces of most notably SIP [2] for session control, Diameter [3] for policy/charging control; plus for PES telephony established protocol interfaces of H.248 [4] for bearer control and Signalling System #7 (SS7) UK Integrated Services User Part (UK-ISUP) [5] for Time Division Multiplex (TDM) interconnect. Whilst mature H.248 with extension package of Notification Rate (NR) [6] protects core from gateways and UK-ISUP with Adaptive Automatic Congestion Control (A-ACC) [7] do provide robust OCP, immature SIP with "Retry-After" and Diameter with "Diameter_Too_Busy" responses that alone are insufficient to provide robust OCP.

The key IMS PES functions and associated components are: Application Server (AS) that performs service logic of calls, Home Subscriber Server (HSS) that provides unified database for user authentication and profile, Call Session Control Function (CSCF) that comprises Serving CSCF (S-CSCF) to determine AS to register to, Interrogating CSCF (I-CSCF) to determine S-CSCF handling user and Proxy CSCF (P-CSCF) that is the first point of contact for SIP controlled end user User Agent (UA) endpoints. For SIP user connections these pass through User Network Interface (UNI) trust boundary security component of Access Session Border Controller (A-SBC) before terminating on the P-CSCF. H.248 controlled end user Access Media GateWay (A-MGW) component endpoints are under control of Access Gateway Control Function (AGCF). H.248 controlled TDM interconnect termination Trunk Media GateWay (T-MGW) component endpoints are under control of Media Gateway Control Function (MGCF). For SIP trunk connections these pass through Network Network Interface (NNI) trust boundary security component of Trunk Session Border Controller (T-SBC) and are handled using Interconnect Breakout Control Function (IBCF). Media Server (MS) component provides announcements, tones and conferencing packet replication capabilities as a network resource – where these are not provided elsewhere such as on A-MGW, T-MGW or UA components. Electronic NUmber Mapping (ENUM) function provides mapping of E.164 format telephone numbers to Universal Resource Identifier (URI). Lastly the Charging Collection Function (CCF) collects usage detail records for call accounting purposes. Depicted in Figure 1 below are IMS PES related key protocols, functions and components.

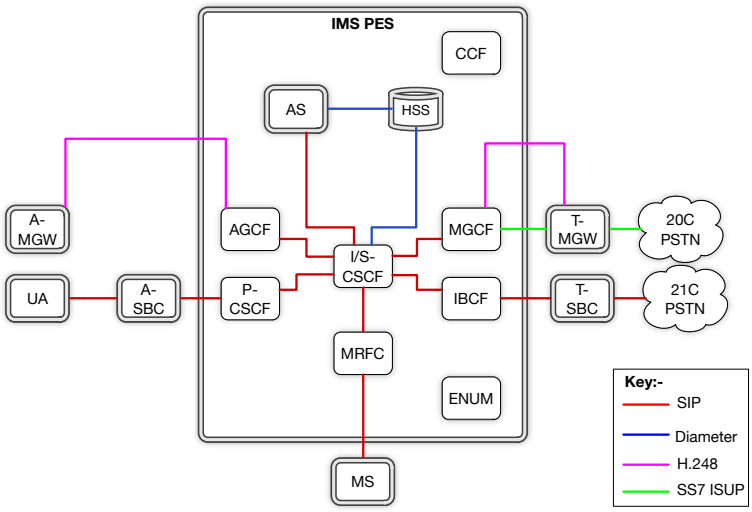

Figure 1: IMS Based NGN Platform for Telephony Service Showing Interface Signalling Protocols

3. Overload.

Overload is operating an entity – eg a node, component, function, etc - beyond its rated capacity. Where this is excessive, ie beyond engineered/economic capacity, then it is likely to result in service performance and/or platform integrity degradation or loss. Degradation can take form of reduced call completion where traffic carried less than traffic offered and may even be below rated capacity, increased transactional latency which can give rise to increased Post Dialling Delay (PDD), etc.

This impact is undesirable given revenue reduction and/or brand damage. Overload as an issue is more serious when it comes to an NGN platform because with an NGN call control is highly centralised; with in extremis a minimum of two nodes serving millions of customers across whole of a country.

Overload can arise for many reasons and these can be categorised as either being planned or unplanned. Planned occurrences include Mass Session Event (MSE) associated with scheduled audience participation televotes such as in UK the Autumn season X-Factor and Strictly Come Dancing television programmes. Unplanned occurrences happen for many reasons and include MSE arising from adverse weather, public transport accident, terrorist act, rioting breakout, etc and Mass Registration Event (MRE) arising from component failure/restoration or code/configuration errors resulting in messaging storms.

To mitigate the impact of an overload it is necessary to apply controls in a timely and decisive manner. The controls can either be Manual or Automatic.

For manual controls Network Operation Centre (NOC) staff will do this; either unilaterally or multilaterally depending on whether the scope is intra- and/or inter-network respectively. The control will typically take the form of some upstream 1-in-'n' gapping, where possible, to reduce the load that the core subjected to. Paramount is early detection through alarm and/or real time monitoring of key resources – eg Processor, Memory and Queue utilisation.

For automatic controls these range from sophisticated and proven capabilities such as A-ACC on UK-ISUP and NR on H.248 to "Diameter_Too_Busy" response on Diameter and "Retry-After" response on SIP that alone are insufficient in that there use is ambiguous – for more than just overload – and do not specify client side behaviour. ETSI for Diameter have recently specified the use of Generic Overload Control Application Protocol (GOCAP) [8]. But not for SIP. There is work going on in Standards Development Organisation (SDO) of Internet Engineering Task Force (IETF) in its SIP Overload Control (SOC) Working Group (WG) but so far, as of IETF Meeting #81 in July 2011, this has not resulted in an Internet Draft (ID) or Request For Comment (RFC) on an implementable mechanism.

It is the latter that is the focus of this paper, ie SIP Overload Control.

4. Test Bed.

The principle objective for the area of research is to evaluate effectiveness of SIP overload control mechanisms. If necessary, to bring about improvements in dynamic behaviour, proposing new/enhanced algorithms, protocols, implementation, etc as needed.

To progress the area of research need to establish a test bed. This comprises TH being developed and that will be used to characterise the efficacy of SUT overload control mechanisms, both current as baseline and new/enhanced ones; at rated load [x1.0], overload [x1.1] and extreme overload [x5.0] rates of SIP registration and session invites.

The TH is based on Personal Computer (PC) technology. It can comprise either single or multiple PCs providing transponder and/or management function. The number of transponders depends on the load to source/sink through the SUT and the Central Processor Unit (CPU) and Network Interface Card (NIC) throughput of the PC(s) being used. The management function is to capture and report telemetry from the transponders and, where possible, from the SUT Man Machine Interface (MMI) also. Statistics accumulated during execution runs includes number of: endpoints registered, calls attempted, calls successful, calls failed, calls concurrently in progress, calls instantaneous throughput, indicative [based on signalling] PDD, etc. Each transponder is tasked with one or more process(es) to simulate Registrant, Caller and/or Callee endpoints. It may be necessary to sequentially apply processes; for instance to register simulated endpoints using Register Process (RP) before starting between simulated endpoints Calling Process (CgP) to corresponding Called Process (CdP). For testing to be representative the traffic characteristics shall be consistent with those observed in the behaviour of real customers, eg say 25 milliErlangs per line residential traffic in each direction and 300 second Mean Call Hold Time (MCHT). Also, given the focus of PATS based on IMS PES, then instantiation of call flows needed for emergency as well as ordinary calling. The way in which the processes are physically/logically connected and the pre-conditioning applied to the SUT depends on the use case being simulated, eg to induce registration storm arising from loss of AGCF resulting in associated AGWs having to failover their association from their primary to secondary AGCF; whilst calling and re-registration associated with AGWs primary on the remaining AGCF continue uninterrupted. Driving SUT through rated load, overload and extreme overload is not onerous to achieve using TH implemented on readily available hardware. The test bed concept is depicted in Figure 2 below.

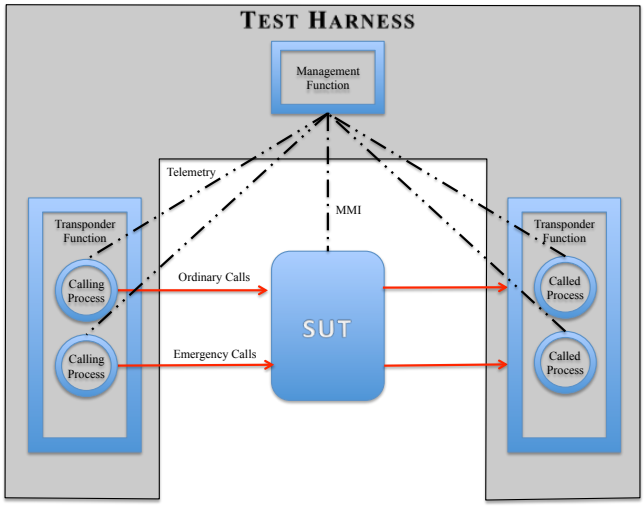

Figure 2: Test Bed to investigate dynamic behaviour of SUT

Furthermore, the test bed will also allow witnessing of the dynamic behaviour of the SUT. Looking for loss of integrity, failing to progress registration and calls, state machine becoming hung, overcorrection by overload control leading to oscillation in effective throughput and increased time to restoration, degraded end user service experience – eg excessive variance in PDD, identify which component/function of an IMS PES constitutes the bottleneck and hence is most in need of overload control on it and associated interfaces, etc.

4. Test Cases.

Proposed test cases will examine:-

1. Current 503 Retry-After Header mechanism to provide benchmark “control”
2. Simple SIP Leaky Bucket algorithm to provide benchmark to see if novel improvements of value
3. Extended SIP Leaky Bucket algorithm prioritised based on SIP Method
4. Extended SIP Leaky Bucket algorithm prioritised based on Calling Party Category being 999 call
5. Extended SIP Leaky Bucket algorithm prioritised based on both 3. and 4. plus possibly
6. Enhancement to apply backpressure to traffic source(s) based on gapped call rate
7. Enhancement to optimise overload control for different use contexts of SIP UNI or NNI and
8. Enhancement to optimise SIP implementation itself for more efficient handling, eg of registration.

Lastly, may also want to factor into consideration an extension to SIP to provide overload control not just that improves for telephony but other multimedia services in general too.

5. Conclusions.

This paper has provided an overview of an NGN platform based on IMS PES; describing the components and interfaces and that the newer protocols used of SIP and Diameter in particular are immature from overload standpoint and do not yet provide robust means to mitigate the impact of Register/Invite storms and this can and does give rise to overload intra- and inter- node. It also has stated what constitutes overload, what is likely to cause overload and what overload control mechanisms are defined on pertinent interfaces in standards and discusses the adequacy of these.

Furthermore, it has gone on to describe a test bed that will be used to characterise the efficacy of both current mechanisms/algorithms for overload control as well as potentially being used to identify and test enhancements to see if brings about demonstrable improvement in performance and/or dynamic behaviour.

Acknowledgments.

This work is supported under Contract 49034 by BskyB and UCL/EPSRC.

References.

[1] ITU-T, “General overview of NGN”, ITU-T Y.2001, 2004-12.

[2] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261 (Proposed Standard), 2002-06, Updated by RFCs 3265, 3853, 4320.

[3] Calhoun, P., Loughney, J., Guttman, E., Zorn, G. and Arkko, J., "Diameter Base Protocol", RFC 3588 (Proposed Standard), 2003-09.

[4] ITU-T, “GateWay Control Protocol”, ITU-T H.248.1, 2005-09.

[5] NICC, “ISDN User Part”, NICC ND1007, 2007-01.

[6] ETSI TISPAN, “NGN Overload Control Architecture; Part 4: Adaptive Control for the MGC”, ETSI ES 283 039-4 V2.1.1, 2007/04.

[7] NICC, “ISUP Overload Controls”, NICC ND1115, 2001-08.

[8] M. Whitehead, “GOCAP - one standardised overload control for next generation networks,” BT Technology Journal, vol. 23, no. 1, pp. 144–153, 2005.

[9] V.Hilt, E. Noel, C. Shen, A. Abdelal, “Design Considerations for SIP Overload Control” IETF SOC WG Internet Draft, 2011-05, Work In Progress.