International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Series H

Supplement 6

(04/2006)

SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS

Control load quantum for decomposed gateways

ITU-T H-series Recommendations – Supplement 6

ITU-T H-SERIES RECOMMENDATIONS

AUDIOVISUAL AND MULTIMEDIA SYSTEMS

CHARACTERISTICS OF VISUAL TELEPHONE SYSTEMS	H.100–H.199
INFRASTRUCTURE OF AUDIOVISUAL SERVICES	
General	H.200–H.219
Transmission multiplexing and synchronization	H.220–H.229
Systems aspects	H.230–H.239
Communication procedures	H.240–H.259
Coding of moving video	H.260–H.279
Related systems aspects	H.280–H.299
Systems and terminal equipment for audiovisual services	H.300–H.349
Directory services architecture for audiovisual and multimedia services	H.350–H.359
Quality of service architecture for audiovisual and multimedia services	H.360–H.369
Supplementary services for multimedia	H.450–H.499
MOBILITY AND COLLABORATION PROCEDURES	
Overview of Mobility and Collaboration, definitions, protocols and procedures	H.500–H.509
Mobility for H-Series multimedia systems and services	H.510–H.519
Mobile multimedia collaboration applications and services	H.520–H.529
Security for mobile multimedia systems and services	H.530–H.539
Security for mobile multimedia collaboration applications and services	H.540–H.549
Mobility interworking procedures	H.550–H.559
Mobile multimedia collaboration inter-working procedures	H.560–H.569
BROADBAND AND TRIPLE-PLAY MULTIMEDIA SERVICES	
Broadband multimedia services over VDSL	H.610–H.619

For further details, please refer to the list of ITU-T Recommendations.

Supplement 6 to ITU-T H-series Recommendations

Control load quantum for decomposed gateways

Summary

This Supplement defines a baseline for control load metrics for H.248 systems with focus on performance engineering parameters relevant for control processing in H.248 network nodes, on correspondent performance design objectives relevant for H.248 network nodes, and on examples of traffic models.

Source

Supplement 6 to ITU-T H-series Recommendations was agreed on 13 April 2006 by ITU-T Study Group 16 (2005-2008).

Keywords

H.248, load control, NGN, performance, traffic model.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this publication, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this publication is voluntary. However, the publication may contain certain mandatory provisions (to ensure e.g. interoperability or applicability) and compliance with the publication is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the publication is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the publication development process.

As of the date of approval of this publication, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this publication. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

		Page
1	Paradigm shift – Motivation	1
1.1	Purpose	1
1.2	Scope and initial objectives	1
1.3	Linearity assumption	2
2	References	2
3	Terminology and definitions	3
3.1	Session versus Call	3
3.2	General definitions	3
3.3	BHxA-related definitions	4
4	Abbreviations	6
4.1	Mathematical symbols	8
5	Basic model for 2-party communication services	9
5.1	Network model	9
5.2	Session variants	10
6	Processing performance	14
6.1	Idealized model	14
6.2	Session processing performance	15
6.3	Context processing performance	15
6.4	H.248 performance classes	16
7	Capacity	19
7.1	Theoretical capacity	19
7.2	Engineered capacity	19
8	Reference Control Load	19
8.1	Session Processor load parameters	20
8.2	Context Processor load parameters	20
9	Session-to-Context relation	21
9.1	Background	21
9.2	1:1 relationship	22
9.3	1:N relationship	23
10	Extensions for the basic control load quantum	25
10.1	Extension factors	25
10.2	Throughput reduction factors	26
10.3	Reduced effective throughput in case of extended H.248 context processing	26
Appendix I – Fundamental relations		26
I.1	Relation between Effective Multiplication Factor κ and Extension Factor *e*	26

Page

Appendix II – Basic traffic models for H.248 systems 27
II.1 Lost context model 27
II.2 Overload Control Model 27
II.3 Combined control/user plane model for H.248 Contexts of type "Circuit-to-X" 32
II.4 Effective throughput versus Context Holding Time: $\phi_{CoCPS} = f(C_O HT)$ 36
II.5 Overload Control Model for access gateways 38
II.6 Overload Control Model for ITU-T Rec. H.248.11 40
Appendix III – Examples of control processing capacity computations 42

Supplement 6 to ITU-T H-series Recommendations

Control load quantum for decomposed gateways

1 Paradigm shift – Motivation

The successful control load quantum in traditional circuit-switched networks (CSN): *Busy Hour Call Attempts* (BHCA), for a time unit 'hour', respectively denoted as *Call Attempts per Second* (CAPS), for a time unit 'second', as well as the corresponding control performance quantum *Busy Hour Call Completions* (BHCC), respectively denoted as *Call Completions per Second* (CCPS), are misleading in H.248 network nodes.

NOTE 1 – "Traditional" refers to the call definition and control load understanding according to ITU-T Rec. Q.543 [4], the control performance framework for digital switching systems. See also ITU-T Rec. Y.1530 [5].

An H.248-based packet-switched network (PSN) is (1) architecturally different in comparison to legacy CSNs, particularly in the following three principal aspects:

- *decomposed control structure* into H.248 MGC and H.248 MG, whereby the main vertical control processing portion is part of the 'controller';
- *server* approach, by centralizing the distributed control of many legacy switching systems into a few number of session control servers; and
- the typical *1:N relation* with regard to the MGC-to-MG ratio.

It is obvious that any reuse of legacy terminology requires a careful handling and common understanding.

NOTE 2 – The reuse of 'BHCA', 'CAPS', etc., is possible in H.248 environments, particularly in the scope of PSTN/N-ISDN service emulation. But it is not recommended, particularly due to potential misunderstandings and the PSTN/ISDN extending scope of H.248.

Additionally, the architectural motivation for the network is based on a technical incentive that requires a "BHCA mapping" on H.248 network nodes: implying that a knowledge of (2) load control and overload protection mechanisms is a prerequisite for understanding the underlying control load quantum. For example, the H.248.11 *Overload Control Package* defines a tight cooperation principle between a MGC and associated MGs; H.248.11 applies the same principles to load quantification.

(3) A third aspect concerns relating the pure Packet-to-Packet (Pa2Pa) MG application with session control protocols at a MGC level, i.e., without the presence of a direct *call* relationship (e.g., 3GPP IP Multimedia Subsystem – IMS).

1.1 Purpose

This Supplement introduces BHC_OA (Busy Hour Context Attempts) as a baseline control load metrics for H.248 systems, and defines a control load quantum based on a basic H.248 context. It includes the definition of performance engineering parameters relevant for control processing in H.248 network nodes and the definition of performance design objectives relevant for H.248 network nodes. This Supplement also provides examples of processing capacity calculations.

1.2 Scope and initial objectives

The objectives of the current edition are:

- identification of the need for an extended performance engineering framework in the context of decomposed control platforms;
- introduction of new terminology (such as BHC_oA, BHSA, effective multiplication factor);

- initial definition of a control processing model;
- initial definition of H.248 Context-based performance classes; and
- basic relations of load and performance parameters according to the defined performance framework.

The initial scope is to achieve consensus on a qualitative basis, the natural next step would be then to commence quantitative performance investigations.

1.3 Linearity assumption

Linearity is assumed. Also, first-order traffic engineering calculations frequently use linearization approximations, particularly in the context of control load estimations (like BHC_aA)[1].

2 References

[1] ITU-T Q-series Suppl. 31 (2000), *Technical Report TRQ.2141.0: Signalling requirements for the support of narrow-band services over broadband transport technologies – Capability Set 2* (CS-2).

[2] *ITU-T Vocabulary: SANCHO Database* (ITU-T Sector Abbreviations and Definitions for a Telecommunications Thesaurus Oriented database), http://www.itu.int/sancho.

[3] ITU-T Recommendation E.600 (1993), *Terms and definitions of traffic engineering*.

[4] ITU-T Recommendation Q.543 (1993), *Digital exchange performance design objectives*.

[5] ITU-T Recommendation Y.1530 (2004), *Call processing performance for voice service in hybrid IP networks*.

[6] VILLAR (J.E.): Traffic Calculations in SPC Systems, *8th ITC*, November 1976.

[7] ITU-T Recommendation E.492 (1996), *Traffic reference period*.

[8] ITU-T Recommendation E.500 (1998), *Traffic intensity measurement principles*.

[9] ITU-T Recommendation E.501 (1997), *Estimation of traffic offered in the network*.

[10] ITU-T Recommendation E.502 (2001), *Traffic measurement requirements for digital telecommunication exchanges*.

[11] ITU-T Recommendation E.503 (1992), *Traffic measurement data analysis*.

[12] ITU-T Recommendation E.508 (1992), *Forecasting new telecommunication services*.

[13] ITU-T Recommendation E.529 (1997), *Network dimensioning using end-to-end GOS objectives*.

[14] ITU-T Recommendation E.711 (1992), *User demand modelling*.

[15] *Generic Requirements for Voice over Packet End-to-End Performance*. Telcordia GR-3059-CORE (March 2000).

[16] *Switching System Overload Control Generic Requirements*. Telcordia TR-NWT-001358, (September 1993).

[17] *LSSGR: Traffic Capacity and Environment*. Telcordia GR-517-CORE (December 1998).

---

[1] For example, [6]: The assumption of a *linear relationship* between processor *occupancy* and *offered load* (BHCA) holds well in *steady-state*, *fault-free* conditions with a *constant call-type distribution*, up to the designed occupancy level for overload capacity.

[18] *ETSI TR 182 015, Architecture for control of processing overload in next generation networks.*

3 Terminology and definitions

3.1 Session versus Call

The telecommunication network specific term "call" is often translated to the term "*session*" for packet-switched connectionless networks (e.g., Internet). The notion of a *session* is also fundamental to IP-based NGN architectures. A **session** extends the traditional notion of a **call** in telecommunication networks. An "H.248 session/call" and the associate creation of an "H.248 Context" is typically triggered by a specific *Call Control Protocol* (e.g., SS7 TUP, SS7 ISUP, BICC, DSS1, H.225/H.245, etc.), or a *Session Control Protocol* (e.g., SIP, SIP-I, SIP-T, NGN-SCP) events. The differentiation between a "call" and a "session" is transparent and is actually not too relevant from an H.248 perspective. Both may be used interchangeably from the Gateway Control Protocol point of view. The key control association is fundamentally the H.248 Context.

NOTE 1 – ITU-T Rec. E.600 [3] defines the individual terms "call", "call attempt", and "busy hour", primarily in the context of BHC_aA (Busy Hour Call Attempts). See also the ITU-T terms and definitions database [2].

NOTE 2 – SIP uses the notions of "call", "session" and "dialog" in different aspects (see IETF documents).

In order to avoid confusion with the legacy BHCA definition, it is recommended that the terms "BHSA" and "BHC_oA" be used in the context of H.248 network nodes. That is the reason why the term 'Session' is continuously used in this Supplement.

3.2 General definitions

3.2.1 session/call: 'Session' or 'Call' is a generic term related to the creation, modification and deletion of an H.248 Context (in a MG). Normally, a qualifier is necessary to make clear the aspect being considered, e.g., session attempt. This definition is aligned with ITU-T Rec. E.600 [3].

3.2.2 session/call attempt: 'Session/Call Attempt' is an attempt to achieve the creation of one or more new H.248 Context(s) in the MG. This definition is aligned with ITU-T Rec. E.600 [3].

3.2.3 load: 'Load' means the total number of the various types of attempts presented to a MGC (e.g., a Call attempt from a PSTN terminal or a Session attempt from a SIP user agent) or a MG (e.g., a Context Attempt by the primary MGC) during a given interval of time (i.e., offered load). This definition is aligned with the performance objectives of ITU-T Rec. Q.543 [4].

3.2.4 session load: See Figure 1.

3.2.5 context load: MG Context Load; see Figure 1.

3.2.6 processor: 'Processor' denotes the logical entity responsible for all control processing work. The technical realization may be very different, from a single CPU to multi-processor systems, in any form of cluster organization (e.g., distributed, hierarchical, load and/or functional sharing modes, etc.).

These definitions are illustrated in Figure 1.

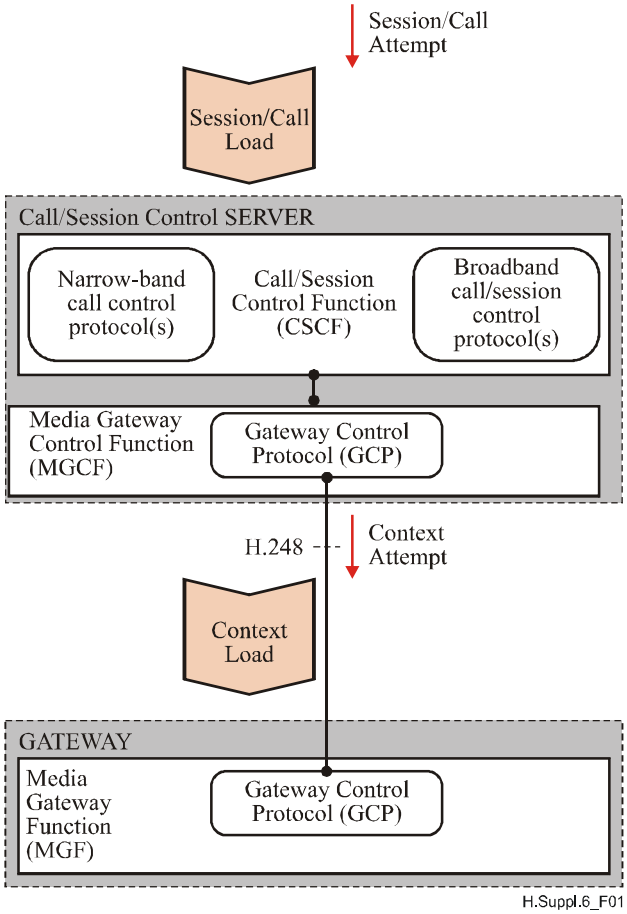

Figure 1 – "Context Attempts" and generated "Context Load"

3.3 BHxA-related definitions

The following table provides a list of generic, BHxA-related load parameters, and corresponding technology-specific example parameters.

BHC_aA $BHC_{Q.543}A$ (shorthand: BHCA)	Busy Hour Call Attempts NOTE – 'call' = PSTN or N-ISDN call according to ITU-T Rec. Q.543.
BHC_bA $BHC_{Q.19XX}A$	Busy Hour Bearer Connection Attempts NOTE – 'bearer connection' = connection controlled by ITU-T Rec. Q.19XX BICC CS1, CS2, CS3 Bearer Control Function (BCF).
BHC_oA $BHC_{H.248}A$	Busy Hour Context Attempts NOTE – 'context' = ITU-T Rec. H.248 Context.

BHC_oA_{MG}	Busy Hour Context Attempts on Media Gateway level NOTE – 'context' = Media Gateway Context for either one of the following H.248-based MG types: – IETF RFC 3525/ITU-T Rec. H.248.1 Media Gateway (MG); – ITU-T Rec. Q.1950 Bearer Interworking Function (BIWF) or Media Gateway Unit (MGU)[a]; – 3GPP 29.232 Circuit-Switched Media Gateway Function (CS-MGW); – 3GPP 29.332 IP Multimedia Media Gateway Function (IM-MGW); – ITU-T "SG 11" Packet Gateway Function (PGF); – ITU-T Rec. J.171.2 Media Gateway (MG)[b].
BHC_oA_{MGC}	Busy Hour Context Attempts on Media Gateway Controller level NOTE – 'context' = Media Gateway Controller Context for either one of following H.248-based MGC types: – IETF RFC 3525/ITU-T Rec. H.248.1 Media Gateway Controller (MGC); – ITU-T Rec. Q.1950 Call Service Function (CSF); – 3GPP 29.232 Mobile Switching Centre Server (MSC Server)[c]; – 3GPP 29.332 Media Gateway Control Function (MGCF); – ITU-T "SG 11" Packet Gateway Control Function (PGCF); – ITU-T Rec. J.171.2 Media Gateway Controller (MGC).
BHSA	Busy Hour Session Attempts
$BHS_{SIP}A$ $BHSA_{RFC3261,SIP}$	Busy Hour Session Attempts NOTE – 'session' = according to IETF RFC 3261 Session Initiation Protocol.
$BHS_{SCP}A$ $BHSA_{NGN\text{-}SCP}$	Busy Hour Session Attempts NOTE – 'session' = according to Draft ITU-T TRQ.ncapx *NGN Session Control Protocol requirements*.
$BHS_{SIP}A$ $BHSA_{3GPP,SIP}$	Busy Hour Session Attempts NOTE – 'session' = according to 3GPP 24.229 IP Multimedia Call Control Protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP).

[a] See TRQ.2141.0 Annex C.

[b] Reference: ITU-T Rec. J.171.2, *IPcablecom Trunking Gateway Control Protocol (TGCP); TGCP Profile 2*, November 2005. The "TGCP Profile 2" is based on ITU-T Rec. H.248, and entitled "TGCP_H248".

[c] E.g., Serving MSC Server, Gateway MSC Server.

NOTE – The difference between Busy Hour Context Attempts on MG level BHC_oA_{MG} and MGC level BHC_oA_{MGC} is illustrated in Figure 10.

The corresponding performance, BHxC-related definitions are appropriate.

Finally, a technical BHxA-related load parameter is provided which is useful for performance considerations on the MG level:

$BHC_{h,DSP}A$	Busy Hour Channel Attempts NOTE – 'channel' = general resource component type "media conversion unit" (MCU) within a MG; a technical realization for a MCU is a "DSP channel"[a]. Note that a "DSP Channel" is the intra-system segment of a user plane connection (e.g., bearer channel) related with a DSP component.
[a] Channel in this sense is the basic "capacity unit" for a digital signal processor in H.248 MG systems. NOTE – The term "mean value" is understood to be the expected value in the probabilistic sense.	

4 Abbreviations

This supplement uses the following abbreviations:

ALN	Analog Line (H.248 Termination physical type)
BHC_aA	Busy Hour Call Attempts
BHC_bA	Busy Hour Bearer Connection Attempts
BHC_hA	Busy Hour Channel Attempts (NOTE – e.g., DSP Channel)
BHC_oA	Busy Hour Context Attempts
BHC_oA_{MG}	Busy Hour Context Attempts (H.248 Context on MG level)
BHC_oA_{MGC}	Busy Hour Context Attempts (H.248 Context on MGC level)
BHSA	Busy Hour Session Attempts
BHSC	Busy Hour Session Completions
BICC	Bearer Independent Call Control
C	H.248 Context
C2C	Circuit-to-circuit (see 5.2.4)
C2P	Circuit-to-packet (see 5.2.2)
C2X	C2X denotes either a C2C or a C2P Session variant
C_aAPS	Call Attempts per Second
C_aCPS, CCPS	Call Completions per Second
C_aHT, CHT	Call Holding Time
C_oAPS	Context Attempts per Second
C_oCPS	Context Completions per Second
C_oHT	Context Holding Time
CP	Context Processor (H.248)
	Control Path (System)
CSCF	Call/Session Control Function
CSN	Circuit-Switched Network (ITU-T Recs H.246, H.332, Y.1001)
DSP	Digital Signal Processor (general)
e	Extension Factor (see 10.1)

FAS	Facility Associated Signalling
GCP	Gateway Control Protocol
IUA	ISDN Q.921 User Adaptation Layer (ITU-T Rec. Q.921, RFC 4233)
MCU	Media Conversion Unit
MEGACOP	Media Gateway Control Protocol (= H.248)
MG	Media Gateway
MGC	Media Gateway Controller
MGCG	Media Gateway Control Function
MGF	Media Gateway Function
MSC	Mobile Switching Centre
NGN	Next-Generation Network
Pa2Pa	Packet-to-Packet
Pe2Pe	Peer-to-Peer NOTE – The abbreviation "P2P" could cause confusion as to whether meaning "peer-to-peer" or "packet-to-packet" and will therefore be avoided in this Supplement.
PSN	Packet-Switched Network
r	Reduction Factor (see 10.2)
SAPS	Session Attempts per Second
SCN	Switched-Circuit Network (ITU-T Rec. H.247) Switched Communication Network (ITU-T Rec. G.177) Signalling Communication Network (ITU-T Rec. G.7712/Y.1703) NOTE – "SCN" and "CSN" denote the same thing in the context of H.248 network nodes. In this Supplement, only the abbreviation "CSN" shall be used, due to the ambiguousness of the abbreviation "SCN".
SCP	Session Control Protocol
SCPS	Session Completions per Second
SG	Signalling Gateway
SHT	Session Holding Time
SIP	Session Initiation Protocol
SP	Session Processor
STM	Synchronous Transfer Mode
TDM	Time Division Multiplexing NOTE – H.248 Termination for *Synchronous Transfer Mode* (STM) interfaces, i.e., TDM is used to abbreviated *Synchronous Time Division Multiplexing* (STDM) [but not *Asynchronous TDM* (ATDM)].

4.1 Mathematical symbols

λ	Arrival rate	[s^{-1}]	Mean arrival rate of service requests[a)]
λ_{CoAPS}	MGC "Context Attempt" rate	[s^{-1}]	Mean "Context Attempt" rate generated by a MGC for a MG
μ	Service rate	[s^{-1}]	Mean service rate of the processing entity[b)]
$\mu_{Context}$	Context Service rate	[s^{-1}]	Mean service rate per H.248 Context
ρ	Utilization		Mean occupancy of a processing entity
ρ_{CcC}	Utilization factor		Mean occupancy of a processing entity by completing H.248 Contexts
ρ_{CcR}	Utilization factor		Mean occupancy of a processing entity by rejecting H.248 Contexts
ϕ	Throughput rate	[s^{-1}]	Mean throughput rate of served requests
$\phi_{Context}$	Throughput rate	[s^{-1}]	Mean effective H.248 Context throughput rate
ϕ_{CoBPS}, ϕ_{CoB}	Context Blocking rate	[s^{-1}]	Mean rate of blocked H.248 Contexts
ϕ_{CoCPS}, ϕ_{CoC}	Context Completion rate	[s^{-1}]	Mean rate of completed H.248 Contexts
ϕ_{CoRPS}, ϕ_{CoR}	Context Rejection rate	[s^{-1}]	Mean rate of rejected H.248 Contexts
h_{Co}, $h_{Context}$	Service time	[s]	Mean service time per H.248 Context
h_{CoC}	Service time	[s]	Mean service time per completed H.248 Context
h_{CoR}	Service time	[s]	Mean service time per rejected H.248 Context
A	Offered load	[Erl]	
A_{CP}	Offered load	[Erl]	Mean offered load per Context Processor
B	Blocking probability		
Y	Carried traffic	[Erl]	
Y_{CP}	Carried traffic	[Erl]	Mean carried traffic per Context Processor
Ω	Queue occupancy		Message buffers, etc.
τ	Delay	[s]	Mean delay of a message

a) E.g., Control plane events: for instance, session initiation messages, call Setup messages, H.248 ADD requests, etc.; User plane events: any type of packet arrivals (e.g., IP packet, MAC frame, ATM cell, AAL2 CPS-Packet, FR frame).

b) Technical realizations: e.g., CPUs, DSPs, IP Forwarding Engine, ATM SAR device, Ethernet switch, etc.

4.1.1 Indices

…$_{Co}$ …$_{Context}$	Context	H.248 Context
…$_{CP}$ …$_{ContextProcessor}$	Context Processor	MGC or MG embedded Context Processor
…$_{CoA}$	Context Attempts	Load
…$_{CoC}$	Completed Contexts	Performance: "Goodput"
…$_{CoR}$	Rejected Contexts	Performance: "Badput" (e.g., rejected, blocked, discarded Contexts)

…BL	Basic Load	Basic (or background) server load, i.e., the none-H.248 related load
…HL	High Load	
…NL	Nominal Load	Engineered capacity, recommended operating point for a considered resource
…OL	Overload	

NOTE – The attribute 'mean' in system/performance parameter notations characterizes the "time mean" (of the underlying stochastic process). However, the purpose of this Supplement is also to provide worst-case estimations for system/performance parameters. These specific requirements will be denoted by an additional index, as follows:

…min	Minimum	Minimum requirement with regard to worst-case assumptions
…max	Maximum	Maximum requirement with regard to worst-case assumptions

5 Basic model for 2-party communication services

The control load quantum shall be based on a basic teleservice, a conversational communication between two session parties.

NOTE – The same principle was applied in PSTN/N-ISDN by using speech telephony service between two calling parties (caller & callee) for "basic call" definitions.

5.1 Network model

The 2-Party property leads to H.248 Contexts types with two H.248 Terminations. H.248 Context processing is done on an MGC and MG level. The scope of this Supplement is beyond the H.248 Context level, and shall comprise session processing as well. The two technical network elements shall be denoted as *session control server*, and *gateway*. Figure 2 shows that simplified architectural network model.

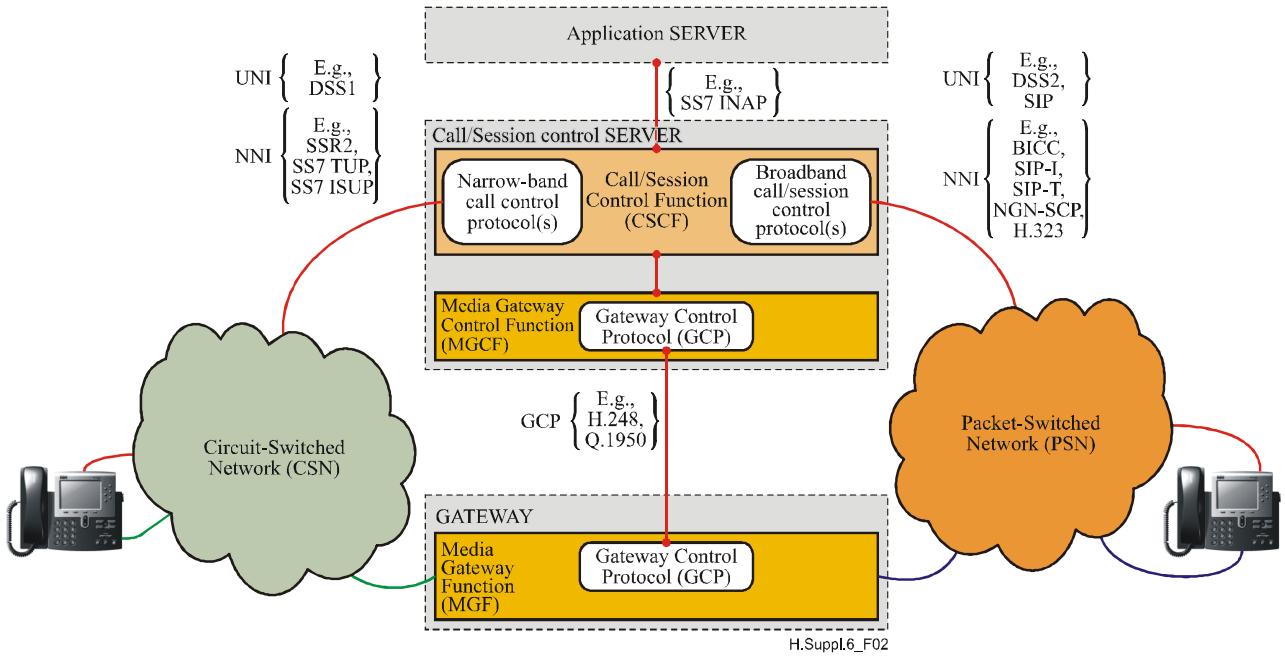

Figure 2 – NGN domains transport, control and application

The dashed boxes are physical network elements (gateway, session control server, application server). The rectangles represent functional entities:

- Media Gateway Function (MGF);
- Media Gateway Control Function (MGCF);
- Call/Session Control Function (CSCF).

NOTE 1 – These functional entities are apparently the most common ones used in various NGN models in ITU-T, 3GPP, ETSI.

The round-cornered rectangles point to what are considered to be the three major generic control protocols: Gateway Control Protocol (GCP), and call/session control protocols for circuit- and packet-switched networks. The double braces show an example of control technologies for the various signalling interfaces. Of course, the specific GCP is H.248, and all other H.248-based control interfaces such as ITU-T Rec. Q.1950, 3GPP 29.232, 3GPP 29.332, etc.

NOTE 2 – Other GCP types like IPDC, MGCP, and ITU-T Rec. J.171 are out of scope.

Out of scope of this performance Supplement is the specific network level (e.g., customer premises equipment domain, access network domain, or core network domain) where the specific H.248 MG may be deployed. Thus, dedicated performance aspects of residential MGs, access MGs, trunking MGs, etc., will not be considered.

Also out of scope are potential differences between mobile or fixed NGNs.

5.2 Session variants

5.2.1 Overview

ITU-T Rec. H.248 distinguishes between two basic Termination types: physical (PHY) and ephemeral (EPH). Figure 3 summarizes the three resulting Context types for 2-party communication services.

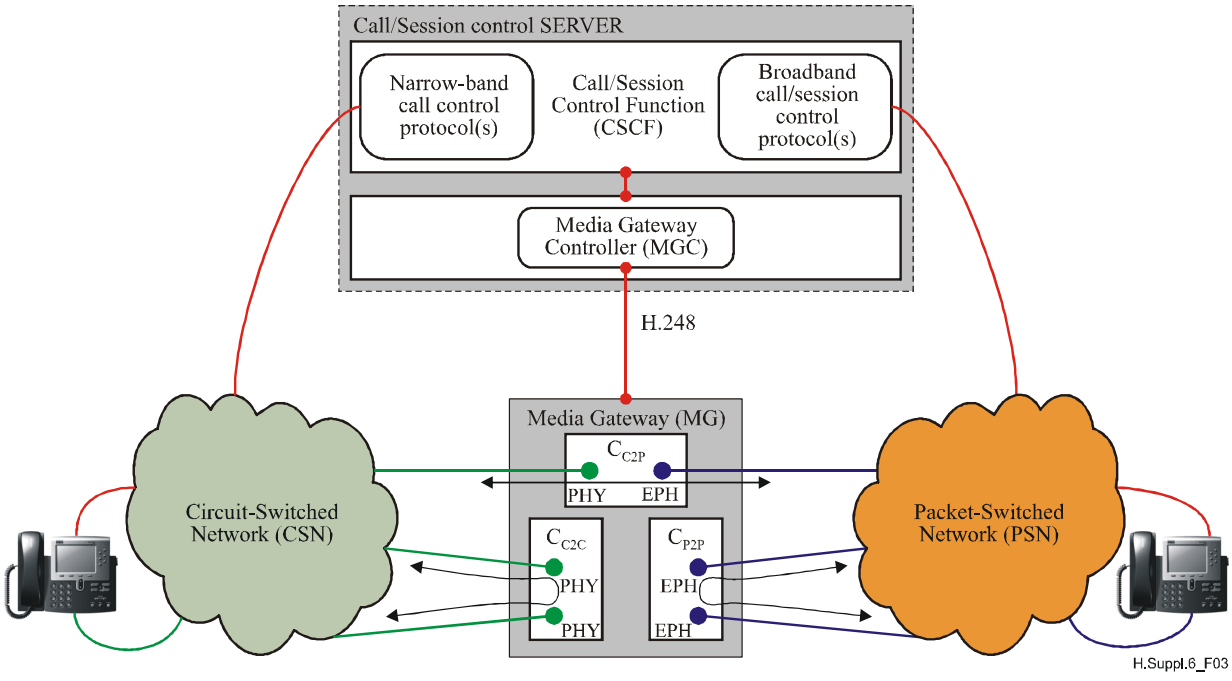

Figure 3 – Session categories – Overview

All three principal Context types represent valid interworking scenarios.

5.2.2 Circuit-to-packet interworking

The circuit-to-packet (C2P) interworking scenario (e.g., voice over Internet Protocol) is the most common one for fixed NGNs. This C2P session type is outlined in Figure 4.

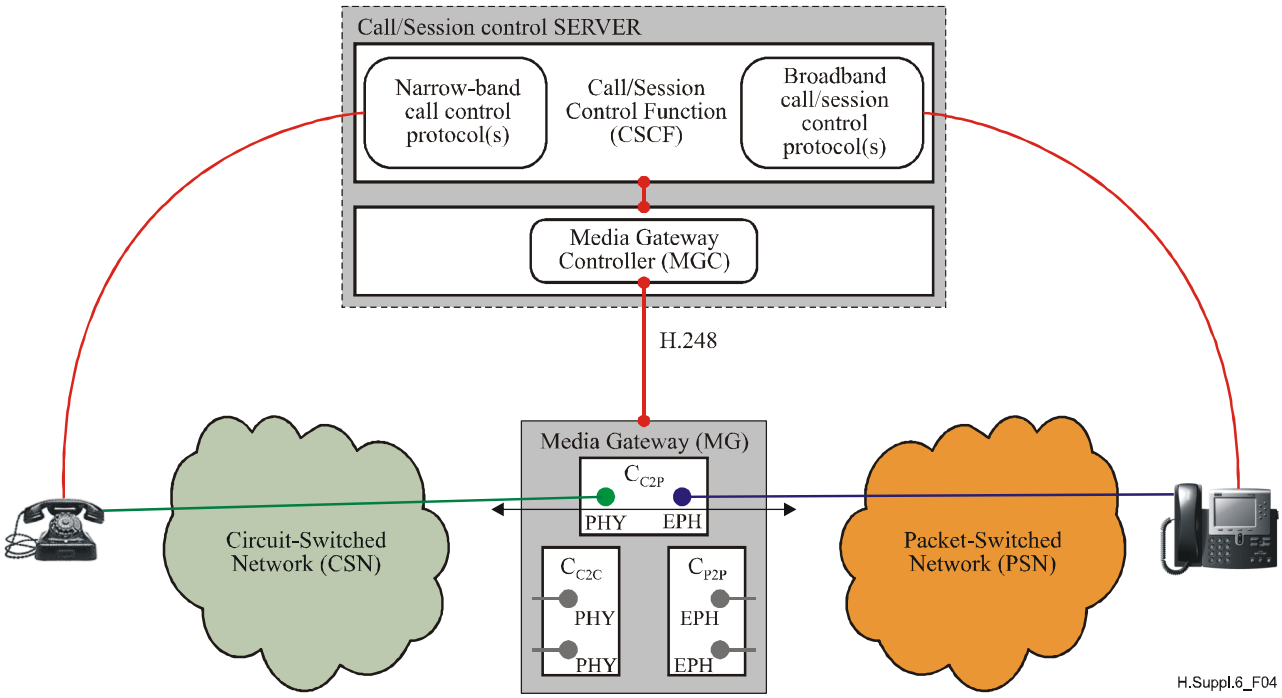

Figure 4 – Session Type (1) – Circuit-to-packet interworking (C2P)

NOTE – The specific H.248 physical Termination type, e.g., TDM for synchronous time division multiplexed interfaces, or ALN for analog lines, is out of scope.

5.2.3 Packet-to-packet interworking

Figure 5 shows the session variant with two ephemeral H.248 Terminations. This interworking case is abbreviated as packet-to-packet (Pa2Pa).

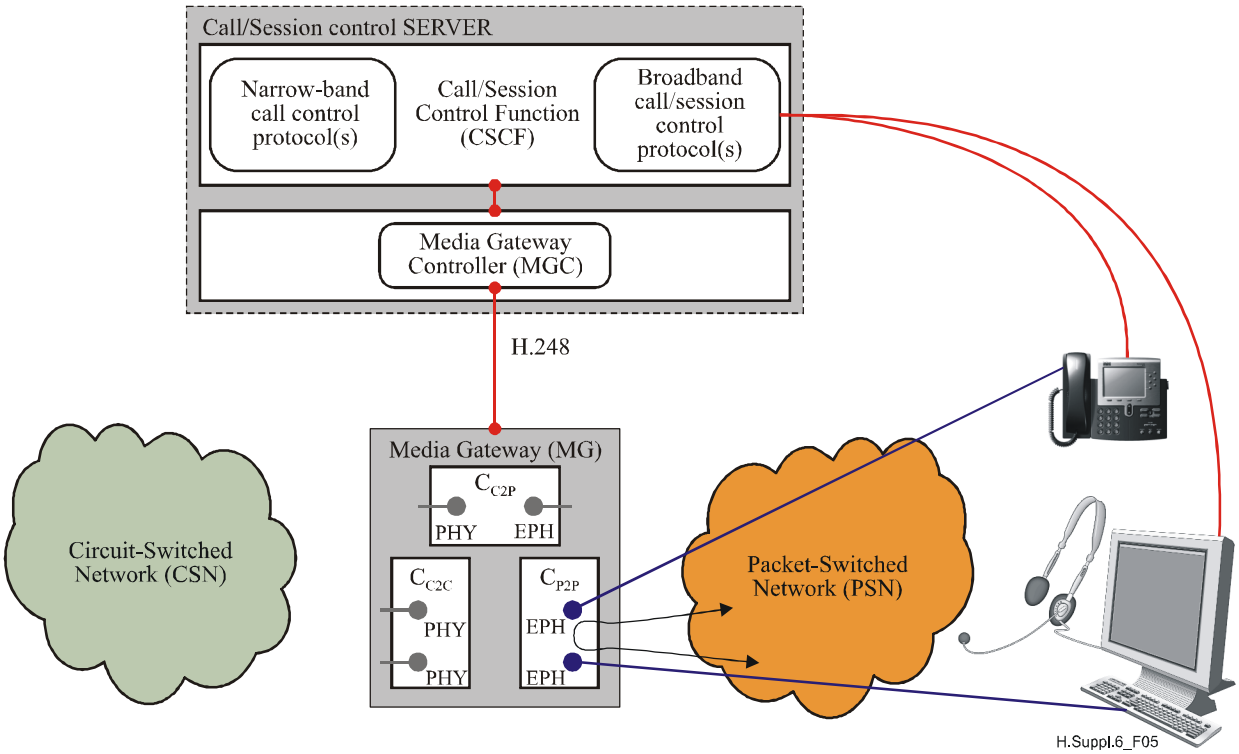

Figure 5 – Session Type (2) – Packet-to-packet (Pa2Pa) interworking

5.2.4 Circuit-to-circuit interworking

The third session variant is circuit-to-circuit interworking (C2C). C2C type sessions are typically needed in order to implement an *internal traffic* type of interworking[2].

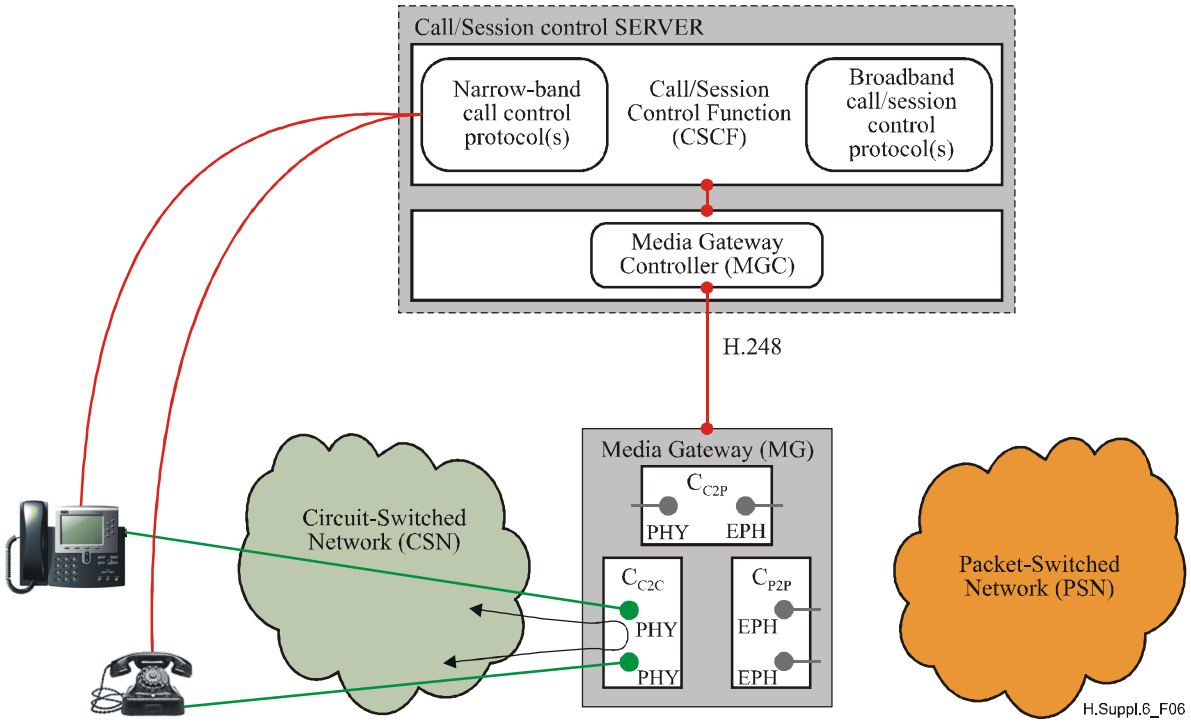

Figure 6 – Session Type (3) – Circuit-to-circuit interworking (C2C)

5.3 Basic H.248 context

A performance framework for control load metrics should be built on H.248 Contexts comprised of two H.248 Terminations. Such a Context shall be denoted as a Basic H.248 Context, similar to the basic call definitions for legacy General Switched Telephone Networks (GSTN), or Intelligent Networks (IN).

NOTE – Examples of ITU-T basic call definitions:

ITU-T Rec. Q.1290: A call between two users that consists of communication only, and does not include additional features.

ITU-T Rec. Q.1300: A call involving exactly two communication entities.

First-order performance evaluations for Basic H.248 Contexts must not take into account detailed information such as:

- Session type;
- H.248 Termination type;
- specific physical respectively ephemeral transport technologies.

More explanations about Basic H.248 Context are found in 6.4.

---

[2] **Internal traffic** is "Traffic originating and terminating within the network considered" (ITU-T Rec. E.600). Internal traffic typically exists at local and transit exchanges. Any "CSN exchange" emulation/simulation scenario using H.248 MGs results in C2C type Contexts. Internal Traffic is emulated/simulated by C2C sessions (e.g., TDM-to-TDM, ALN-to-TDM, ALN-to-ALN) in NGNs. Internal traffic corresponds to *Intrasystem Calls* (see Figure 6-1 of GR-517-CORE).

6 Processing performance

Consider the vertical hierarchy of control interfaces in Figure 2, where there are multiple chained instances with different control processing performance requirements. A simplified architecture is proposed in the following.

NOTE – A more detailed view is for instance outlined in ITU-T TRQ.2141.1 Figure 5-2, showing an object reference model for BICC CS2 Call Bearer Control.

6.1 Idealized model

The *monolithic control* of existing TDM switching systems was decomposed by the transition towards NGN architecture. The major control entities considered are:

– the Session Control Processor (briefly, Session Processor), located in the control path of the "session control server" network element, and

– the Context Control Processor (briefly, Context Processor), located in the control path of the "gateway" network element.

Figure 7 shows that simplified *two-level control* hierarchy as an evolution of monolithic controls. This model may be further detailed, e.g., by differentiating the CSCF and MGCF control parts within the session control server.

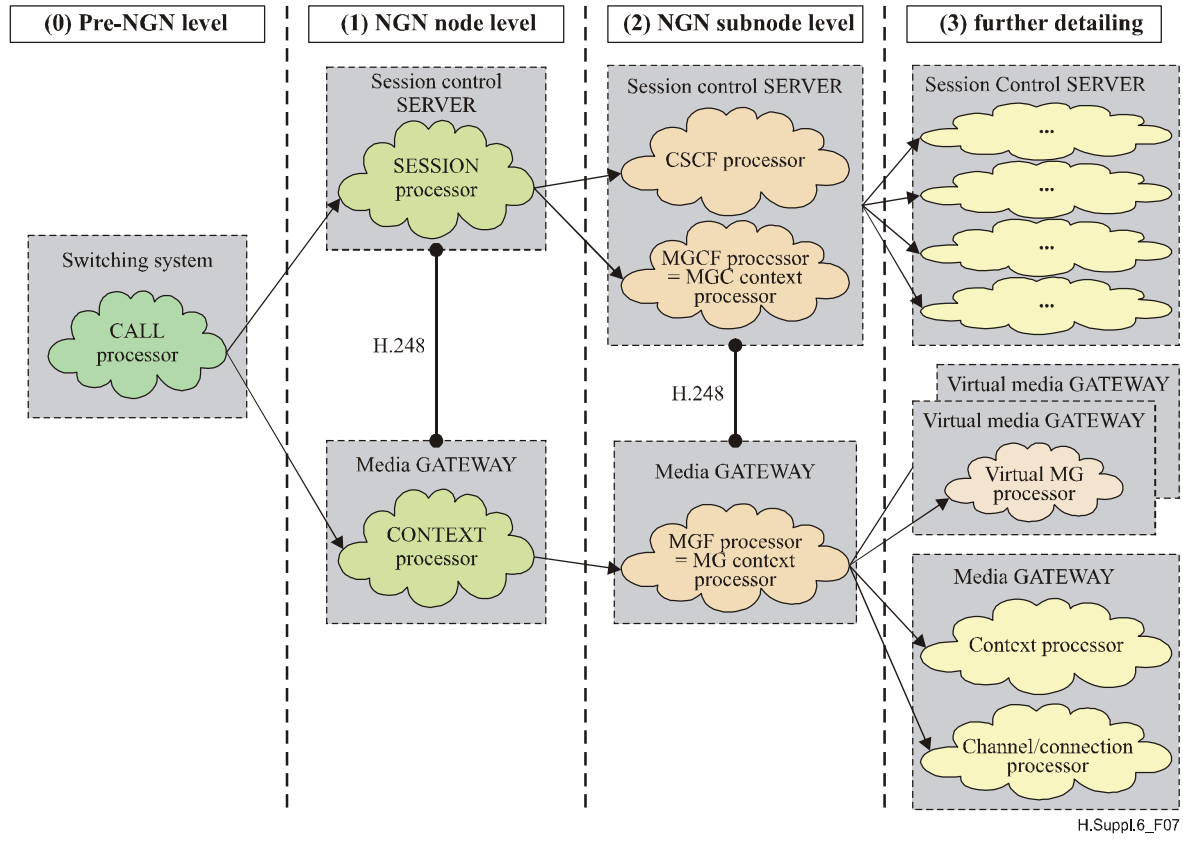

Figure 7 – Control processing model – Potential levels of detailing

The scope of this Supplement is thus indicated in 1) *NGN node level* in Figure 7. Other potential levels are for further study.

NOTE – Potential concepts for refined detailing of MG controls are indicated in Figure 7. The technical motivation behind might be:

a) high-capacity MGs;

b) Virtual MG support; and/or

c) MG-embedded Bearer Control Units, as in the so-called BIWN (Bearer Interworking Node) case in Figure C.2 of ITU-T TRQ.2141.0 [1].

Out of scope are the so-called "combined gateways" because of the existing monolithic style of the control processor, and of the absence of an H.248 interface. Combined gateways are for instance: H.323 gateways, BICC CS1 interworking nodes, 3GPP Release 3 MSCs, or SIP gateways[3] with integrated user and control plane endpoints.

6.2 Session processing performance

Session processing performance is for further study, as the initial scope of this Supplement is the gateway node.

6.3 Context processing performance

The major *performance parameter* is the *effective throughput* figure of merit (sometimes called *goodput*)[4]. Scope is the Media Gateway embedded *Context Processor*. The average service time (in seconds) $h_{Context,Basic}$ for processing *elementary H.248 Contexts* is denoted as indicated in Equation 1.

Average service time per basic H.248 Context $h_{Context,Basic}$

$$h_{Context,Basic} \quad [\mathrm{s}] \tag{1}$$

NOTE – A high-level definition for *Basic H.248 Contexts* was introduced in 5.3. Further discussion is provided in 6.4.

The Ideal Context processor **capacity** (see 7.1 for explanation) is defined by Equation 2, whereas the Ideal **throughput** under ideal conditions is defined by Equation 3.

Context Processor – Maximum service rate $\mu_{Context,Basic}$

$$\mu_{Context,Basic} = \frac{1}{h_{Context,Basic}} \quad [\mathrm{s}^{-1}] \tag{2}$$

Context Processor – Effective context throughput $\phi_{Context,Basic}$ *under ideal conditions*

$$\phi_{Context,Basic} = \mu_{Context,Basic} \quad [\mathrm{s}^{-1}] \tag{3}$$

Equation 3 shows that the stationary throughput is equal to the service rate of the control processor.

6.3.1 Completion rate C_OCPS

Effective throughput for a real context processor under ideal conditions, i.e., every context attempt may be successfully processed according to Equation 4:

Context Processor – Context Completions per second, ϕ_{CoCPS}

$$\phi_{CoCPS} = \phi_{Context,Basic} \quad [\mathrm{s}^{-1}] \tag{4}$$

---

[3] For example, a SIP gateway housing RTP endpoints together with SIP user agent functionality, as well as for instance CSN circuits together with CSN call control.

[4] The complementary figure, the *ineffective throughput* is often denoted as *badput*. This non-effective throughtput is generating *blind load* in the control processor.

NOTE – 'Ideal' means that every H.248 Context may be successfully served. There are no unsuccessful sessions, error situations, rejected Context requests, inadequately handled Contexts[5], or other cases.

6.3.2 Completion rate BHC_OC

The context completion rate is defined by Equation 5 and given in time unit 'hour^{-1}':

Context Processor – Busy Hour Context Completions, ϕ_{BHCoC} (per hour)

$$\phi_{BHCoC} = \phi_{CoCPS} \cdot 3600 \quad [h^{-1}] \tag{5}$$

6.4 H.248 performance classes

Any meaningful NGN service requires, from an H.248 point of view, at least a single H.248 Context. A 2-party communication service demands a Context with two H.248 Terminations at a minimum. Such a generic Context shall be denoted as a 'Basic Context' (see also 5.3). The necessary control processing performance during the whole lifetime of a Basic H.248 Context shall be associated with a performance class (i.e., Class 1 in Figure 8).

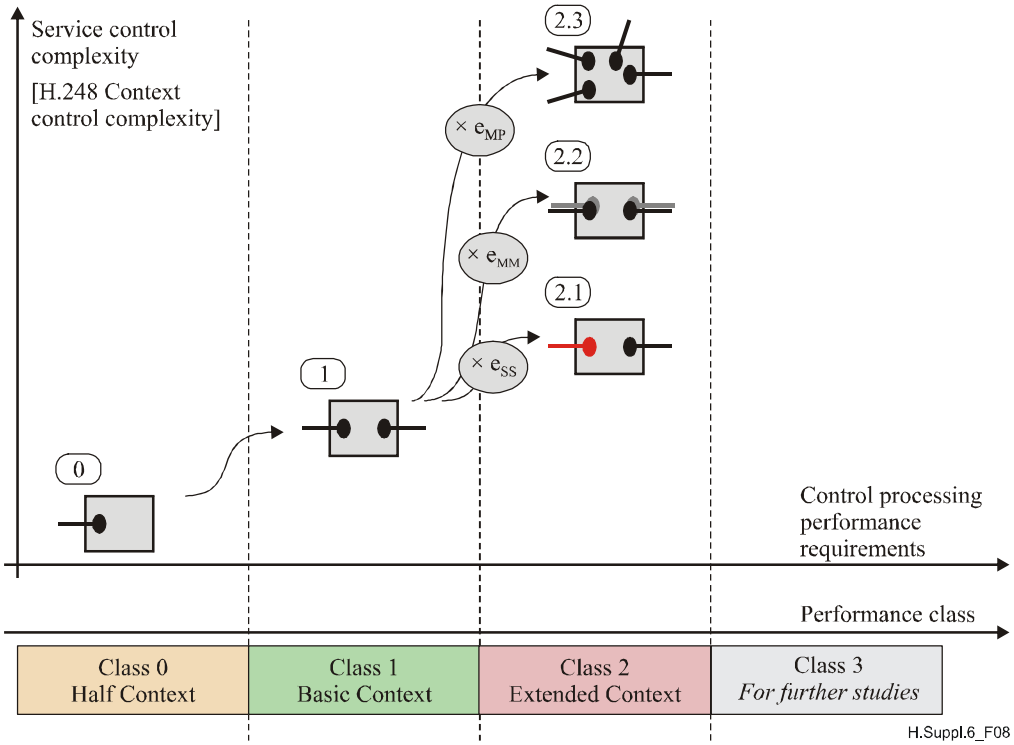

Figure 8 – Performance classes – Qualitative categorization

The principle of differentiating a *basic service* from *extended services*, as, for example, supplementary services, is well known in telecommunication networks. This rule is also applied in performance engineering as the first principle classification for separating basic load requirements and basic performance requirements from additional demands associated with extended services.

NOTE 1 – An "extended service" may be for instance (ITU-T Rec. Q.1741.1) a service which modifies or supplements a basic (telecommunication) service. Consequently, it cannot be offered to a user as a

---

5 "Inadequately handled H.248 Context attempts" can be defined according to ITU-T Rec. Q.543: "[…] are attempts which are blocked (as defined in the E.600-series Recommendations) or are excessively delayed within the MG (or MGC). 'Excessive delays' are those that are greater than three times the '0.95 probability of not exceeding' values recommended in …".

stand-alone service. It must be offered together, or in association with, a basic (telecommunication) service. The same supplementary service may be common to a number of basic (telecommunication) services.

The same principle may be applied in defining separated categories for *Basic H.248 Contexts* and *Extended Contexts*. Figure 8 illustrates such an abstraction concept by various performance classes. From a performance engineering point of view, the Extended Context types will be linked with Basic Context by so-called **extension factors** $e_{(+)}$. Examples of *Extended Context* types will be introduced in the following clause and quantitative dependencies are discussed in 10.1.

NOTE 2 – Performance considerations related to operations on the H.248 Root Termination (e.g., specific audits) are for further study.

6.4.1 Reduced performance necessity

There are processing requirements below the Basic Context level. This is indicated by the "Half Context" case in Figure 8 (Class 0). A control load quantum below the basic level may make sense to cover, for instance, in the following cases:

- abandoned session during establishment phase;
- test signal sequences (e.g., some selected ITU-T Rec. H.248.17 scenarios);
- channel associated signalling (with later Context change);
- digit collection (with later Context change);
- delivery of PSTN supplementary services in on-hook state; or
- others.

NOTE – Whether or not the H.248 Termination of the "half context" belongs to the H.248 Null Context is not to be distinguished.

6.4.2 Potential extension areas

Table 1 provides three initial categories for potential extension areas. The resulting Extended Contexts have extended performance requirements.

Table 1 – Examples of extended Contexts

Class 'Extended'	Extension factor $e_{(+)}$	Class labelling
2.1	e_{SS}	Superset Services (SS) Covers extension from basic services towards additional services *per H.248 Termination*. Examples are inband signalling, channel associated signalling, Subscriber Line Protocol-based PSTN supplementary services, overload protection, etc.
2.2	e_{MM}	Multimedia (MM) Covers extension from monomedia towards multimedia sessions. Examples are single media stream per H.248 Termination, i.e., multiple Terminations per session party; or multiplexed cases: multiplexed media streams, cascaded multiplexing Terminations, etc.
2.3	e_{MP}	Multiparty (MP) Covers extension from 2-Party (2PY) to 3-Party (3PY), and general Multiparty session configurations
2.4		For further study

NOTE – This initial categorization scheme of Table 1 might be too coarse for specific performance engineering cases. A more detailed classification, e.g., by separating e_{SS} in for instance $e_{SS,CAS}$, $e_{SS,CLIP}$, or $e_{SS,Test}$ within class 2.1, is for further study.

The "Extended Class" is one case where there are increased performance requirements *per session*. It shall be noted that another case may be the Session-to-Context ratio (see clause 9 'Session-to-Context relation').

6.4.3 Classification tools

6.4.3.1 Signalling scenario

Signalling scenarios (also known as *Message Sequence Charts*) are often used as first-order qualifiers for the indication of underlying service control complexity. Additionally, second-order qualifiers might be for example the respective *signalling message types*. Particular signalling message *information elements* may act as third-order qualifiers.

A similar approach may be applied for H.248 signalling as well, by considering for instance the mean number of H.248 Commands per session, Context manipulation functions, Termination modifications, etc. An H.248 Context control complexity indicator might be then "derived" from a signalling complexity.

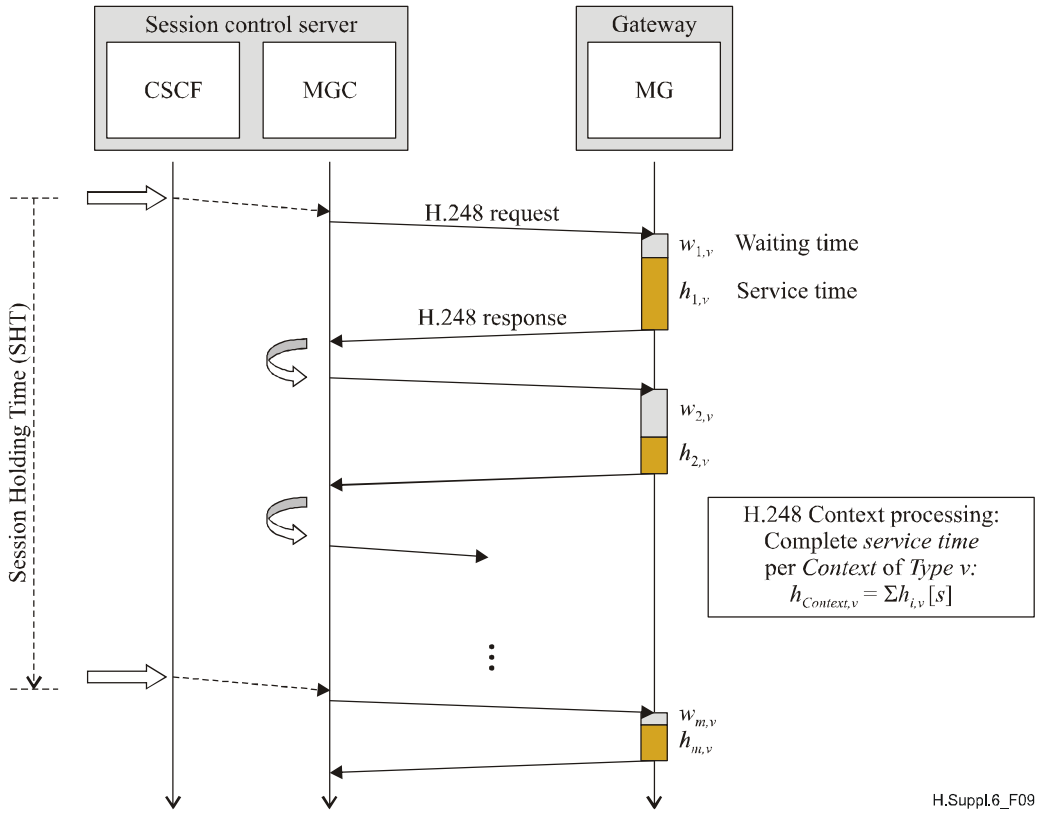

Figure 9 – Generic H.248 signalling scenario

Figure 9 illustrates a generic H.248 signalling scenario. The usage of H.248 signalling scenarios for the derivation of H.248 performance metrics is for further study.

6.4.3.2 Session/Context state machine models

Refined BHC_aA models were often based on the consideration of advanced finite state machines for call modelling. The same principle may be applied for H.248 Context modelling. A Context state machine model approach is for further study.

NOTE – A simple state machine for modelling an H.248 Context lifetime would be to use two Context states, either 'idle' or 'active'. The active state is reached by a Context creation, and left for example by final Termination SUBtract. There may be two further types of state transitions defined for characterizing active-to-active state transitions:

a) MODification events (triggered by MGC); and

b) NOTification events (triggered by MG local events).

Corresponding traffic parameters, e.g., modification rate, notification rate, etc., may be defined for profiling service and thus, for qualifying H.248 performance classes.

6.4.3.3 Code Count method

The Code Count method is a traditional instrument for first-order estimations of performance requirements. This reverse engineering approach is based on the analysis of control software. In the meantime, modern source code analyser tools[6] allow the automatic generation of a variety of software metrics. Some of these metrics might be used for performance classification, e.g., the specific volume metric "number of lines containing source code".

NOTE – Of course, an absolute classification is not possible due to the implementation-specific character of software (e.g., programming language, architecture). However, a relative classification with regard to quantitative categorization of performance classes, as well as the separation of subclasses within a dedicated class, is possible and straightforward.

7 Capacity

Performance is always limited in every technical system by its inherent available capacity. The control processor *capacity* figure is consequently an important link between *performance* (clause 6) and *load* (clause 8). These principles still apply in the case of H.248 network nodes. The main purpose of this clause is to recall the two major capacity terms.

7.1 Theoretical capacity

The theoretical control processing capacity is the maximum service rate, i.e., the maximum session completion rate, which in the H.248 environment is the maximum H.248 Context completion rate. See for instance $\mu_{Context,Basic}$ (in Equation 2) for Basic H.248 Contexts processed by the Context processor.

7.2 Engineered capacity

The engineered capacity is always below the theoretical processor capacity. If a Session/Context-based definition is required in future, then a Q.543-based adaptation is recommended.

NOTE – ITU-T Rec. Q.543 "Engineered Capacity": The mean offered load at which the exchange just meets all the **grade of service requirements** used by the Administration to engineer the exchange.

8 Reference Control Load

The purpose of this clause is to focus on the *load parameters* related to Context processing. Further *Performance* objectives (in addition to those indicated in clause 6) are scoped in the subsequent clauses. Figure 10 shows the main dependencies between the several load factors and corresponding performance types. The control processing model is based on the "*NGN subnode level*" as shown in Figure 7.

---

6 For instance: www.scitools.com, …

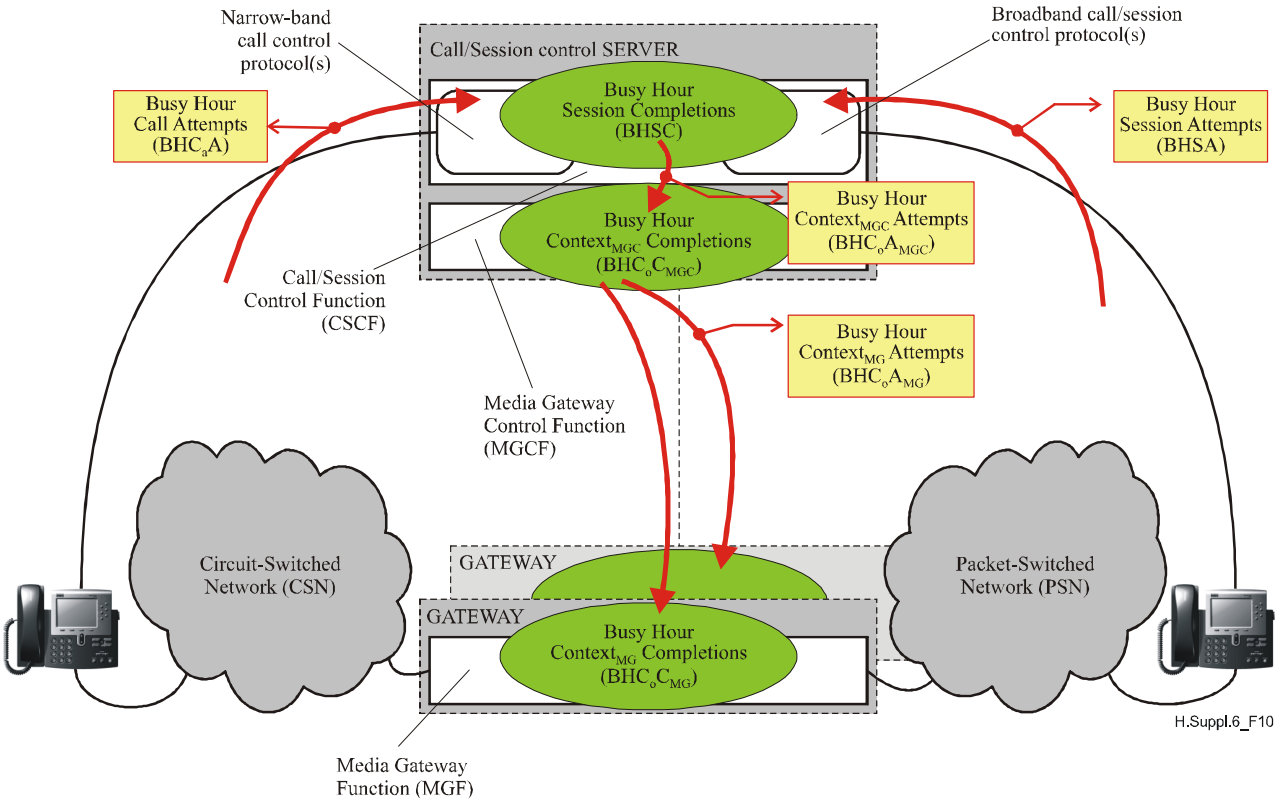

Figure 10 – Control Processing Model – Load/Performance chaining

NOTE 1 – Even though load is sometimes equated with performance, this definitely is not the general case for telecommunication systems like H.248 nodes. Of course, under certain conditions, performance equals the load. For instance, in stationary *low load* situations, BHC_oC may be estimated using the BHC_oA value.

NOTE 2 – A simple model describing the principal load-performance behaviour may be the *Lost Context Model*, see clauses II.1 and II.2.3.

The two-processor model continues to be assumed in the following clauses (as illustrated in Figure 7 *NGN Node Level*).

8.1 Session Processor load parameters

The arrival rate of session attempts may be defined in seconds and hours time unit levels.

8.1.1 Arrival rate SAPS

The rate of session attempts per second is denoted by λ_{SAPS} (Equation 6).

Session Processor – Session attempts per Second, λ_{SAPS}

$$\lambda_{SAPS} \quad [s^{-1}] \tag{6}$$

8.1.2 Arrival rate BHSA

The session attempt rate in time unit 'hour^{-1}' is defined in Equation 7.

Session Processor – Busy Hour Session Attempts, λ_{BHSA} *(per hour)*

$$\lambda_{BHSA} = \lambda_{SAPS} \cdot 3600 \quad [h^{-1}] \tag{7}$$

8.2 Context Processor load parameters

The arrival rate of H.248 Context Attempts may be defined in seconds and hours time unit levels.

8.2.1 Arrival rate C_OAPS

The rate of Context Attempts per second is denoted by λ_{CoAPS} (Equation 8).

Context Processor – Context Attempts per Second, λ_{CoAPS}

$$\lambda_{CoAPS} \quad [s^{-1}] \tag{8}$$

8.2.2 Arrival rate BHC_OA

The Context Attempt rate in time unit 'hour^{-1}' is given in Equation 9.

Context Processor – Busy Hour Context Attempts, λ_{BHCoA} *(per hour)*

$$\lambda_{BHCoA} = \lambda_{CoAPS} \cdot 3600 \quad [h^{-1}] \tag{9}$$

8.2.3 Basic Context Control Load

The **offered load** $A_{ContextProcessor}$ (A_{CP}) to the MG-embedded Context processor, generated by incoming attempts for basic H.248 Contexts, is defined by Equation 10.

Offered load $A_{ContextProcessor}$ *for basic H.248 Contexts*

$$A_{ContextProcessor} = \lambda_{CoAPS} \cdot h_{Context,Basic} \quad [Erl] \tag{10}$$

NOTE 1 – An "incoming attempt" relates to the first H.248 ADD.request command from the MGC for a new H.248 Context.

NOTE 2 – The offered load A_{CP} defined by Equation 10 corresponds to ITU-T Rec. E.500 [8] parameter traffic intensity A [Erl]. Clause 5.2/E.500 describes "traffic intensity concept and stationarity". This E.500 description may be reused by replacing "job" with "H.248 Context", and "resource holding time" with "Context holding type (C_OHT)".

8.2.3.1 Normal load

The definition of a "*Normal Basic Context Control Load*" parameter is for further study. A definition based on ITU-T Rec. E.500 *Normal Load Traffic Intensity* will be recommended (if required in the future).

8.2.3.2 High load

The definition of a "*High Basic Context Control Load*" parameter is for further study. A definition based on ITU-T Rec. E.500 *High Load Traffic Intensity* will be recommended (if required in future).

8.2.3.3 Reference load definitions

Reference load definitions, e.g., for performance class "Basic H.248 Context", are for further study.

NOTE – Telcordia GR-517-CORE [17], or ITU-T Rec. Q.543 [4] provide reference load definitions for digital exchanges. The reference loads are defined by using load parameter types "traffic intensity", "arrival rate", and/or "holding time".

9 Session-to-Context relation

9.1 Background

The H.248 decomposed gateway principle leads to the fact that the correlation between a user plane connection (here *H.248 Context*) and a respective control plane association (here *Session*) disappears from a Media Gateway perspective. The knowledge about the session identifier and corresponding Context identifier(s) is located in the session control server (housing the MGC instance) and the MG does not have that kind of information.

NOTE 1 – The same situation applies for MG-embedded signalling gateways (SG), like IETF SIGTRAN SGs. For instance, in the case of a SIGTRAN IUA SG, the MG does not have the knowledge whether control plane connections (here ITU-T Recs Q.931/Q.921) are associated with user plane connections (here a H.248 Context).

This means that the MG is not able to correlate session control load with Context control load.

NOTE 2 – For instance, in Figure 11 the MG does not know firstly, that the H.248 Contexts $C_{i,j}$ belongs to Session S_i, and secondly, that the two consecutive H.248 Contexts $C_{i,1}$ and $C_{i,2}$ belong to the same Session S_i.

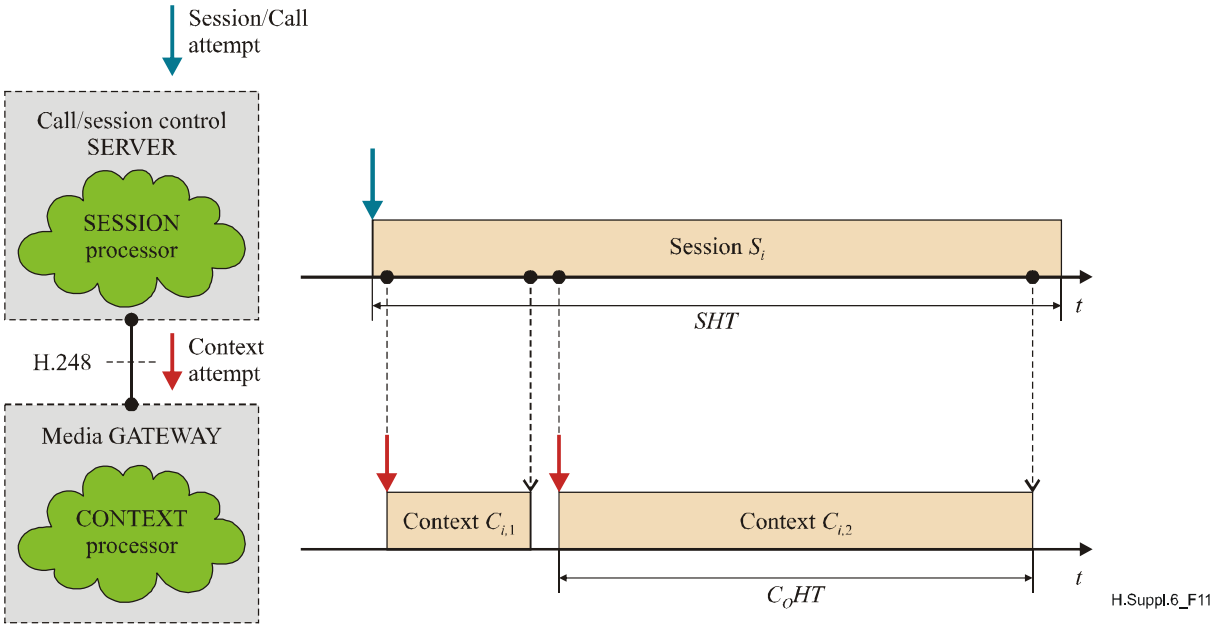

Figure 11 – General Session-to-Context relation

NOTE 3 – The sketched holding times in Figure 11 refer to the *mean Session Holding Time* (SHT) respectively *mean H.248 Context Holding Time* (C_OHT).

9.2 1:1 relationship

There is a 1:1 relationship between a session and a corresponding H.248 Context for the majority of services. This means that in a *1:1 session type* a single H.248 Context C_i must be processed in a Media Gateway behind a single session S_i in the control server.

NOTE – It has to be noted that multiple MGs may be involved in the same session, and all these MGs are controlled by the same session control server. But this does not change the 1:1 relationship from the MG point of view.

Example 1: MGC responsible for one MG in a session

There will be one H.248 Context to be controlled from the MGC side. The Context Attempts rate $\lambda_{CoAPS,MGC}$ will be equal to the Session Attempts rate λ_{SAPS} (when all session attempts are accepted).

Example 2: MGC responsible for two (or more) MGs in the same session

If a MGC controls multiple MGs and the session requires multiple MGs, then there may be for one session multiple context attempts, e.g., one for each MG. The Context Attempts rate $\lambda_{CoAPS,MGC}$ will be at least twice the Session Attempts rate λ_{SAPS} (when all session attempts are accepted).

The Context Attempts rate $\lambda_{CoAPS,MG}$ from the MG perspective is independent of the example scenario.

9.2.1 Control Load – Session or Context arrival rates

The resulting arrival rates on the Session Processor and on the Context Processor level are identical, as indicated in Equation 11.

Arrival rates for 1:1 relationships (per second and per hour)

$$\begin{aligned} \lambda_{CoAPS} &= \lambda_{SAPS} \quad [s^{-1}] \\ \lambda_{BHCoA} &= \lambda_{BHSA} \quad [h^{-1}] \end{aligned} \tag{11}$$

NOTE – Of course, identical arrival rates may not lead to identical load factors on the Session Processor and on the Context Processor. Rather, the usual case is that $A_{ContextProcessor}$ differs from $A_{SessionProcessor}$ due to the *server* approach, i.e., typically is $A_{SessionProcessor} < A_{ContextProcessor}$.

9.3 1:N relationship

There are many services with a 1:N ratio of a single session to the associated number of Contexts in a MG.

An example of a 1:N session type would be session-triggered bearer connection tests before the end-to-end conversation phase using SS7 Continuity Checks for the call/session associated circuit. Such a test might be done via a first H.248 Contexts $C_{i,1}$; the consecutively following conversation is handled by second Context $C_{i,2}$. It shall be noted again that the MG may not correlate both Contexts $C_{i,1}$ and $C_{i,2}$. Other examples are given in 6.4.1.

9.3.1 Rate multiplication factor N

The resulting Context Attempt arrival rate is N times higher than the session arrival rate, as defined in Equation 12.

Arrival rates for 1:N relationships (per second and per hour)

$$\begin{aligned} \lambda_{CoAPS} &= N \cdot \lambda_{SAPS} \quad [s^{-1}] \\ \lambda_{BHCoA} &= N \cdot \lambda_{BHSA} \quad [h^{-1}] \end{aligned} \tag{12}$$

There is typically a mix of 1:1 and 1:N types of sessions in a real network, i.e., the average rate multiplication factor is between 1 and N. The crucial point is that the Context arrival rate is greater than or equal to the session arrival rate (e.g., $BHC_OA \geq BHSA$). Figure 12 illustrates the overall qualitative relationship between Session and Context arrival rates.

NOTE 1 – The Context arrival rate BHC_OA is often used as a load indicator (beside others) for the Context processor local overload protection mechanisms. If 1:N types exist in an H.248 network, then the MG should be cautious in using the BHC_OA parameter in control loops for load regulation, or overload control, due to its lack of knowledge of the real multiplication factor.

NOTE 2 – While the rate multiplication factor N is of type Integer, the average rate multiplication factor $\overline{N}$ is typically a non-Integer type.

NOTE 3 – The resulting average rate multiplication factor $\overline{N}$ leads to a **virtual session attempt rate** (or **virtual call attempt rate**) of $\lambda'_{SAPS,MG} = \overline{N} \cdot \lambda_{SAPS}$ from a H.248 Media Gateway perspective.

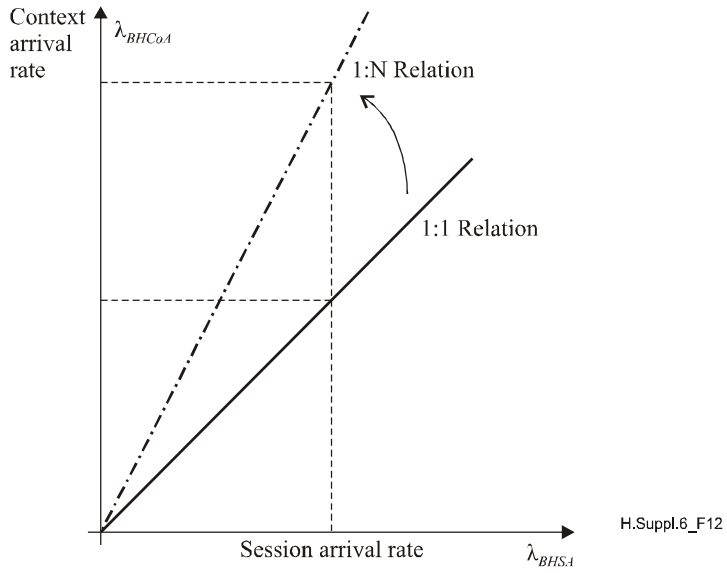

Figure 12 – Session-to-Context proportion – Multiplication factor N between arrival rates

9.3.2 Effective multiplication factor κ

The individual Contexts $C_{i,j}$ may be of different complexity type (see 6.4, H.248 Performance Classes), resulting in different individual mean service times $h_{Context,Ci,j}$ from the Context processor point of view. An *effective multiplication factor* κ characterizes the increased Context processing performance requirements behind a single session (in 1:N session type scenarios). See Equation 13.

Effective multiplication factor κ *based on basic H.248 Context service time* $h_{Context,Basic}$

$$\kappa = \frac{\sum_{j=1}^{N} h_{Context,Ci,j}}{h_{Context,Basic}} \tag{13}$$

NOTE 1 – The effective multiplication factor κ is typically applied as a first-order performance estimation.

Figure 13 illustrates how the increased Context processor load A_{CP} relates to the effective multiplication factor κ.

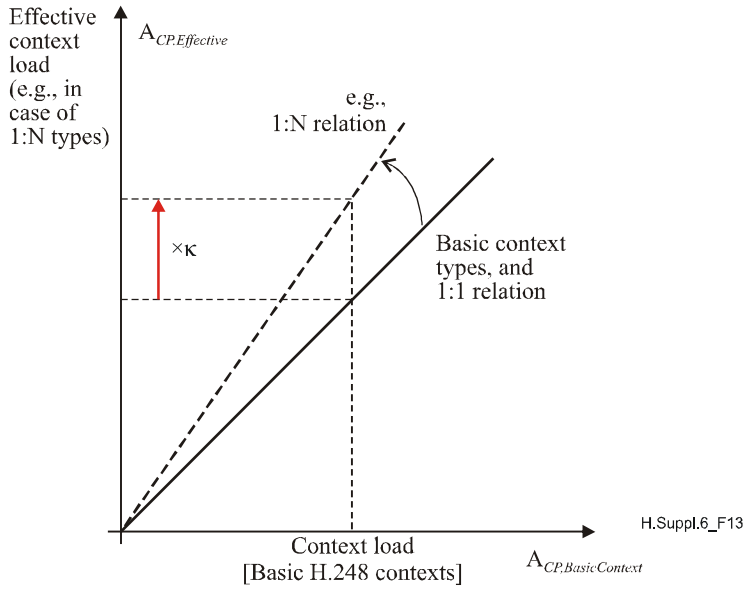

Figure 13 – Context Processor Load A_{CP} – Effective multiplication factor κ

NOTE 2 – For example, the effective Context processor load $A_{CP,Effective}$, applicable on a 1:N Session-to-Context types, may be related to basic Context processing, and estimated with:

$$A_{CP,Effective} = \kappa \cdot A_{CP,BasicContext}$$

10 Extensions for the basic control load quantum

The purpose of this clause is to introduce the additional parameters that are needed for handling the "Extended Context" performance class.

10.1 Extension factors

The additionally needed average service time $h_{Context,(+)}$ extends the required context processor service time as shown in Equation 14.

Average service time per extended H.248 Context $h_{Context,Ext}$.

$$h_{Context,Ext.} = h_{Context,Basic} + h_{Context,(+)} \quad [s] \tag{14}$$

NOTE – The '(+)' is a placeholder for one of the potential extension reasons mentioned in 6.4.2.

A generic extension factor $e_{(+)}$ related to the basic Context service time is introduced by Equation 15.

Generic extension factor $e_{(+)}$

$$e_{(+)} = \frac{h_{Context,Ext.}}{h_{Context,Basic}} = 1 + \frac{h_{Context,(+)}}{h_{Context,Basic}} \tag{15}$$

Equation 16 gives an example for a specific extension factor, e.g., an average figure e_{SS} for class 2 superset services (e.g., PSTN supplementary services).

Examples of specific extension factor e_{SS}

$$e_{SS} = 1 + \frac{h_{Context,SS}}{h_{Context,Basic}} \tag{16}$$

10.2 Throughput reduction factors

The increased service time requirements for extended H.248 Contexts lead to a reduction of the Context completion rate. The generic reduction factor $r_{(+)}$ is shown by Equation 17:

Generic reduction factor $r_{(+)}$

$$r_{(+)} = \frac{1}{e_{(+)}} = \frac{h_{Context,Basic}}{h_{Context,Basic} + h_{Context,(+)}} \tag{17}$$

10.3 Reduced effective throughput in case of extended H.248 context processing

10.3.1 Completion rate $BHC_{O,Ext}C$

The Context completion rate is reduced in comparison to the basic Context completion rate as defined by Equation 18.

Context Processor – Reduced Busy Hour Context Completions $\phi_{BHCo,extC}$ *(per hour) for extended Context processing*

$$\phi_{BHCo,extC} = r_{(+)} \cdot \phi_{BHCoC} \quad [h^{-1}] \tag{18}$$

NOTE – It should be pointed out that rather than the Context Processor performance being reduced, it stays the same, e.g., in terms of program instructions per second performance unit.

Appendix I

Fundamental relations

I.1 Relation between Effective Multiplication Factor κ and Extension Factor *e*

Equation I-1 is derived from Equations 13 and 15 and shows the link between the two linear factors *Effective Multiplication Factor* κ and *Extension Factor e*.

Effective multiplication factor κ *as sum of the individual extension factors* $e_{(+),j}$

$$\kappa = \sum_{j=1}^{N} e_{(+),j} \tag{I-1}$$

Equation I-1 allows a quick first-order load/performance estimation in the case of the known individual class-specific extension factors.

NOTE – The inclusion of class mixes, subclasses, weighting factors, etc., is for further study.

Appendix II

Basic traffic models for H.248 systems

Some basic traffic models for H.248 network nodes are presented for the following performance evaluation areas:

- Lost Context Model (see II.1);
- Basic Overload Control Model for single network nodes (see II.2);
- Overload Control Model for Access Gateways (see II.5);
- Combined Control/User Plane Model (see II.3);
- Control Performance versus Context Holding Time (see II.4).

II.1 Lost context model

Annex B/E.501 [9], *Equivalent Traffic Offered,* describes the basic load-performance dependency in the case of a loss model. The model represents a conservation law. This E.501 "lost call model" can be mapped on to the MG level Context Processor:

In the *lost Context model,* the equivalent traffic offered corresponds to the traffic which produces the observed carried traffic in accordance with Equation II-1.

Lost Context model for H.248 MG Context Processor

$$Y_{CP} = A_{CP} \cdot (1 - B_{CP}) \quad \text{[Erl]} \tag{II-1}$$

where:

Y: is the carried traffic (i.e., *completed Contexts*)

A: is the equivalent traffic offered (*see Equation 10*)

B: is the *Context* congestion through the part of the network (i.e., *MG*) considered

NOTE 1 – This is a purely mathematical concept. Physically, it is only possible to detect "offered traffic" whose effect on occupancies tells whether these attempts give rise to very brief seizures or to calls.

NOTE 2 – The equivalent traffic offered, which is greater than the traffic carried and, therefore, greater than the effective traffic, is greater than the traffic offered when the subscriber is very persistent.

NOTE 3 – B is evaluated on a purely mathematical basis, so that it is possible to establish a direct relationship between the traffic carried and call congestion B and to dispense with the role of the equivalent traffic offered A.

II.2 Overload Control Model

There is an H.248 Context Control Processor at MGC and MG levels (see Figure 7). ITU-T Rec. H.248.11 describes an overload control framework, comprising Context Processors on both MGC and MG levels. While ITU-T Rec. H.248.11 specifies a cooperation principle between MGC and associated MGs realized by a distributed control loop, this clause defines a basic model for local overload controls. "Local" means that the scope of the control loop is spatially limited on the network node, or geographically limited on network node locations.

II.2.1 Theoretical Throughput Model

Figure II.1 shows a single server model for an H.248 Context Processor. The server has two phases. The server is either in idle state, or in phase 'C' in case of successful Context processing, or in phase 'R' in case of rejecting Context Attempts.

NOTE 1 – The target of the Context rejection phase is a protocol conform feedback to the "served user" instance. This is either a call/session control server internal application on top of the MGC in case of an

"MGC Context Processor", or the MGC itself in case of an "MG level Context Processor". The protocol conform reaction shall prevent "repeated Context Attempts".

The H.248 message buffer has a limited size. Fully filled buffers may lead to H.248 traffic loss. The resulting traffic rate shall be denoted as blocked Contexts, to differentiate from the rejection rate.

NOTE 2 – The difference between 'blocking' and 'rejection' is the fact that blocking does not need any server processing time.

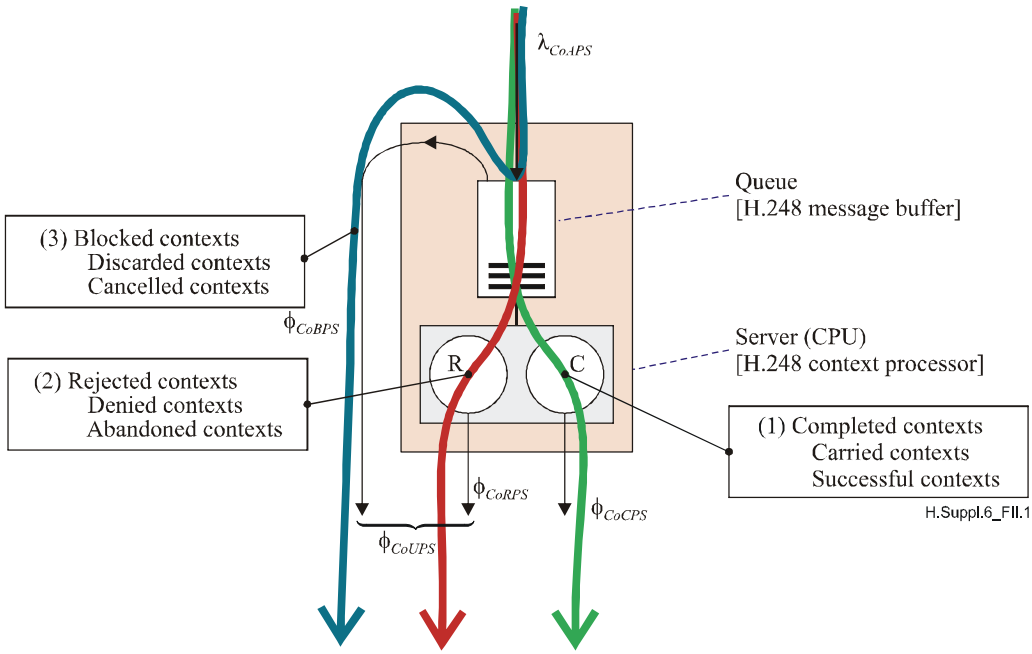

Figure II.1 – Traffic model for ideal throughput considerations

II.2.2 Traffic model for real systems

The queue blocking effect shall be not considered from this point onwards in the text. A real Context Processor is only aware of an H.248 protocol message, if the message is identified as such. Such a protocol analysis is always coupled with processing time. The resulting traffic model is illustrated in Figure II.2.

Every Context Attempt is either successfully handled as completed Context, or rejected.

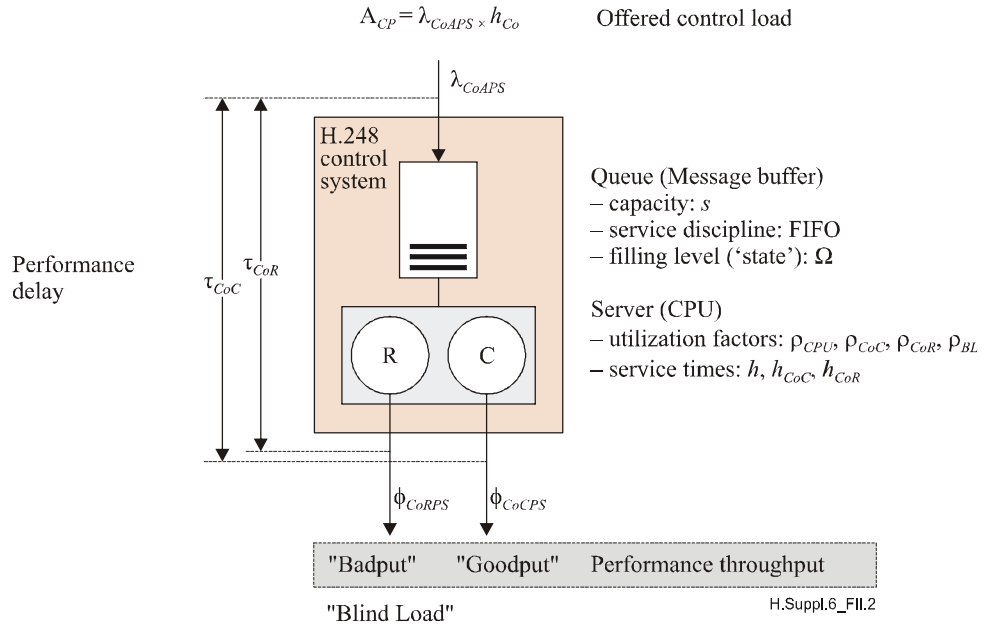

Figure II.2 – Traffic model for overload considerations

It is obvious that the completion of an H.248 Context consumes much more processing time as any unsuccessful Context handling (see also Equation II-3). The system time τ results from service time h_{Co} and waiting times.

II.2.3 Flow analysis

The conservation law is valid under **stationary** conditions; see Equation II-2.

Conservation Law – Stationary context rates

$$\phi_{CoCPS} = \lambda_{CoAPS} - \phi_{CoRPS} \quad [s^{-1}] \tag{II-2}$$

NOTE – Equation II-1 from the Lost Context Model is the dimensionless (in Erl) counterpart to the rate (in s^{-1}) proportions in Equation II-2.

II.2.4 Assumptions

II.2.4.1 Process types

The stochastic arrival and service processes are assumed to have Markov process properties. The traffic model belongs therefore to the class of M/M/1 types. An infinite queue is assumed for later qualitative estimations.

II.2.4.2 Service times

Equation II-3 expresses the fact that unsuccessfully processed or uncompleted H.248 Contexts typically demand less system resources than Context completion.

Qualitative relationship between service times h_{CoR} and h_{CoC}

$$\begin{aligned} h_{CoR} &= \kappa \cdot h_{CoC} \\ h_{CoR} &<< h_{CoC} \end{aligned} \tag{II-3}$$

NOTE – For first-order quantitative estimates a factor κ of 10% may be assumed.

II.2.5 Main Context Processor behaviour

The average *Context serving time* h_{Co} defined in Equation II-4 depends on the stationary operating point ("equilibrium"), and the corresponding *Context completion rate* ϕ_{CoC} and *rejection rate* ϕ_{CoR}.

Average service time per Context $h_{Context}$ *as a function of the operating point*

$$h_{Co} = f(h_{CoC}, h_{CoR}) \tag{II-4}$$

This model and assumptions result in a stationary server behaviour, which is very well-known from conventional Synchronous Transfer Mode (STM) switches (see ITU-T Rec. Q.543 [4]). Figure II.3 illustrates the server utilization factors versus Context Attempt arrival rate.

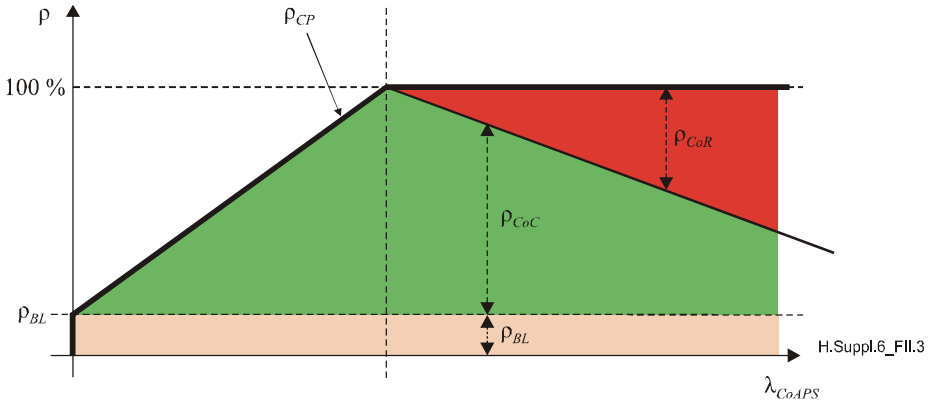

Figure II.3 – Idealized Context Processor behaviour – Server utilization factors versus context arrival rate

II.2.6 Server operation modes – Workload areas for a context processor

The operation mode of an H.248 Context Processor is determined by the Context attempt arrival rate λ_{CoAPS}. Three main server states can be distinguished, as shown by Equation II-5:

Server state – Workload areas dependent on arrival rate λ_{CoAPS}

$$\text{Server}_{\text{State}} = \begin{cases} \text{Underloaded} & 0 \le \lambda_{\text{CoAPS}} \le \lambda_{\text{CoAPS,100\%}} \\ \text{Overloaded} & \lambda_{\text{CoAPS,100\%}} \le \lambda_{\text{CoAPS}} \le \lambda_{\text{CoAPS,Unstable}} \\ \text{Unstable} & \lambda_{\text{CoAPS,Unstable}} \le \lambda_{\text{CoAPS}} \end{cases} \tag{II-5}$$

II.2.6.1 Operation mode "Underload"

Equation II-6 provides the right-hand limit for underloaded server.

Underloaded server – Right-hand limit $\lambda_{CA,100\%}$

$$\lambda_{CA,100\%} = \frac{1-(\rho_{BL}+\rho_{HR})}{h_{CC}} \tag{II-6}$$

II.2.6.2 Operation mode "Overload"

Equation II-7 provides the right-hand limit for an overloaded server.

Overloaded server – Right-hand limit $\lambda_{CA,Unstable}$

$$\lambda_{CA,Unstable} = \frac{1-(\rho_{BL}+\rho_{HR})}{\kappa \cdot h_{CC}} = \frac{1-(\rho_{BL}+\rho_{HR})}{h_{RC}} \tag{II-7}$$

For the limiting operating point, $\lambda_{CA,Unstable}$ is $\phi_{CC} = 0$ and hence $\phi_{RC} = \lambda_{CA} = \lambda_{CA,Unstable}$.

II.2.6.3 Operation mode "Unstable"

Specific metrics for the "unstable" area are not derived.

II.2.7 Throughput estimation

The effective throughput versus control load function $\phi_{CoCPS} = f(\lambda_{CoAPS})$ results in three straight-line equations:

Context Processor operation modes – Straight-line equation $\phi_{CoCPS} = f(\lambda_{CoAPS})$

$$\phi_{CoCPS} = f(\lambda_{CoAPS}) = \begin{cases} \lambda_{CoAPS} & 0 \le \lambda_{CoAPS} \le \lambda_{CoAPS,100\%} \quad \text{Underloaded Server} \\ \dfrac{1-(\rho_{BL}+\rho_{HR})}{(1-\kappa)h_{CoC}} - \dfrac{\kappa}{1-\kappa}\lambda_{CoAPS} & \lambda_{CoAPS,100\%} \le \lambda_{CoAPS} \le \lambda_{CoAPS,\text{Unstable}} \quad \text{Overloaded Server} \\ 0 & \lambda_{CoAPS,\text{Unstable}} \le \lambda_{CoAPS} \quad \text{Unstable Server} \end{cases} \tag{II-8}$$

NOTE – Servers should not be engineered for 100% utilization. There should still be reserves (also known as headroom) under high-load situations. For such, Context Processor reserves are covered by factor ρ_{HR} in Equation II-8.

Figure II.4 summarizes the *goodput function* (top) and *server utilization* (bottom) for the three different workload areas.

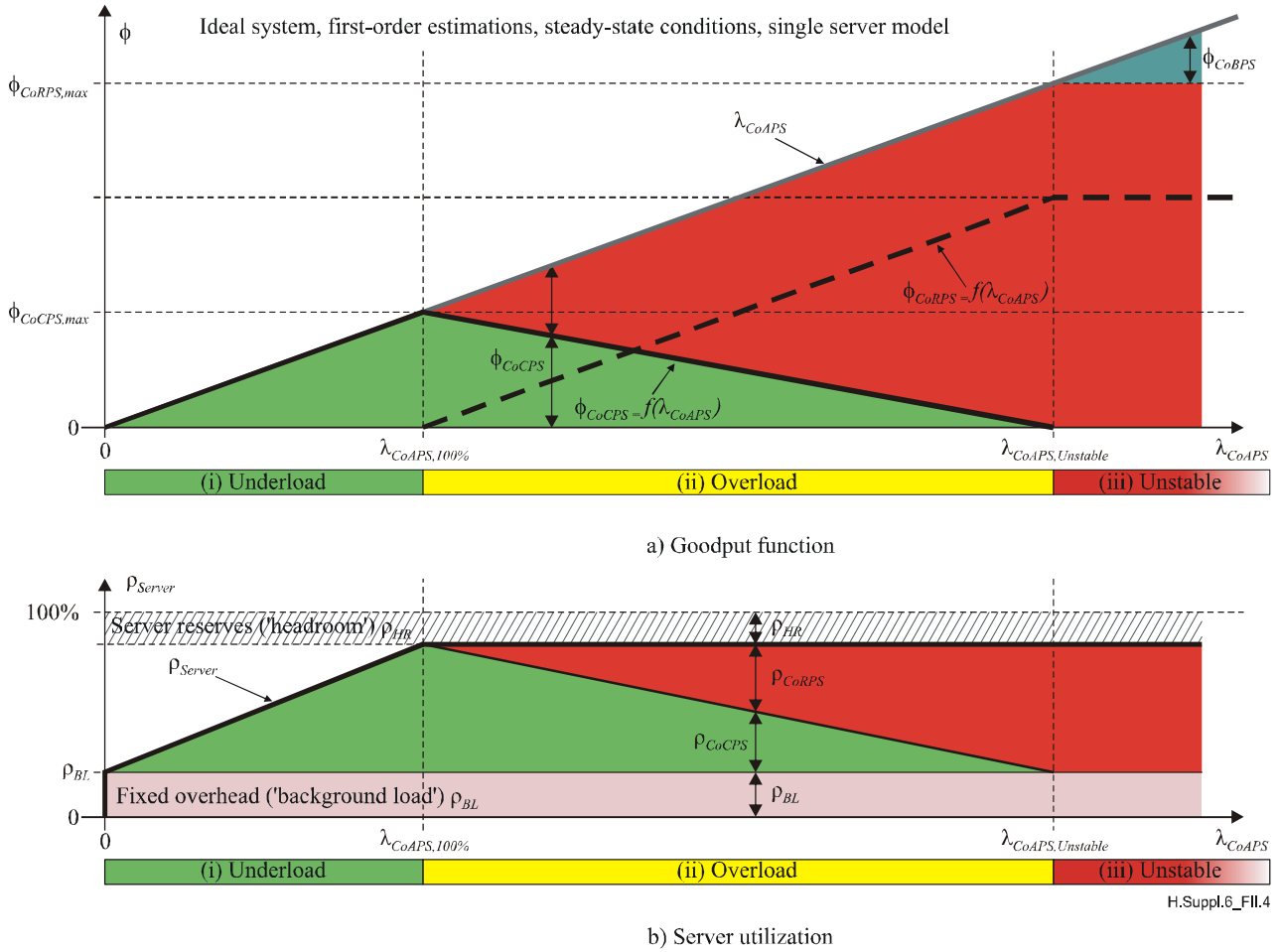

Figure II.4 – H.248 Context Processor operation modes – Goodput and server utilization for the three principal workload areas

II.2.8 Conclusions

This overload control model allows distinguishing three main operation modes of an H.248 Context Processor. Linearization, for first-order estimations, is possible within each operation state. It should be noted that the overall server behaviour is very non-linear.

The maximum Context throughput or goodput$_{max}$ $\phi_{CoCPS,max}$ is:

Optimal goodput $\phi_{CoCPS,max}$

$$\phi_{CoCPS,\max} = \phi_{CoCPS}(\lambda_{CoAPS,100\%}) = \frac{1-(\rho_{BL}+\rho_{HR})}{h_{CoC}} \qquad \text{(II-9)}$$

II.3 Combined control/user plane model for H.248 Contexts of type "Circuit-to-X"

A simple estimation model is presented for a specific class of H.248 Context types.

II.3.1 Background from circuit-switched networks

There is a 1:1 relationship between a call and a bearer connection in circuit-switched networks (CSN). An analog line (ALN), or a TDM circuit, is directly associated with the controlling call. Such a tight coupling leads in the H.248 model to the fact that certain traffic parameters behind a physical H.248 Termination may be easily combined with control plane parameters. This relationship is helpful for engineering H.248 network nodes in case of C2X Context types. Here, C2X denotes either a Session variant C2P defined in 5.2.2, or a Session variant C2C defined in 5.2.4.

II.3.2 Traffic model

Figure II.5 shows an example of a combined user/control plane model for an H.248 Media Gateway. The control path shall be modelled by the *single server* model presented in II.2.2. The server entity is the H.248 *Context Processor* (CP). The MG data path shall be modeled by a *K-server*. The server entity is a *Media Processor* (MP) consisting of *K Media Conversion Units* (MCU). A Media Conversion Unit is responsible for the majority of functions required for service and network interworking.

NOTE 1 – The following terminology will be used: *User plane* and *control plane* are used for external system interfaces, for instance, DS0/E1/PDH as U-plane interface respectively H.248 as C-plane interface. The terms *data path* and *control path* are the respective internal system interface equivalents.

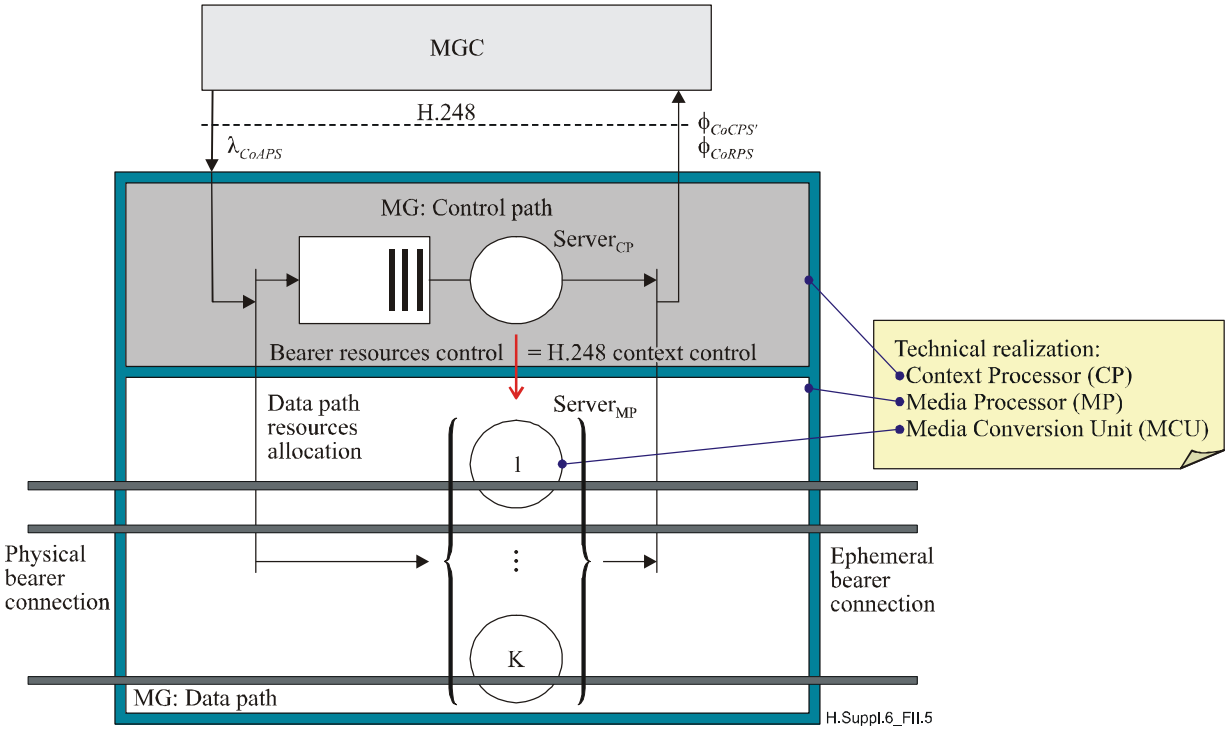

Figure II.5 – Traffic model for H.248 MGs with scope on C2X-type sessions

The control path model is of type *waiting system*, allowing delayed access of H.248 traffic from MGC to get the MG Context Processor resource. The data path model is of type *loss system*; either there is still a free physical H.248 Termination, or all circuits are occupied (in case of C2X-type H.248 Contexts).

The forking element at the ingress side will point out that a new H.248 Context Attempt is internally mapped by the MG on two service requests: one for the Context Control Processor and another for the Media Processor, respectively ("a successful CSN call immediately needs a circuit").

The synchronization element at the egress side is related to the fact that a completed H.248 Context event leads to the simultaneous de-allocation of the corresponding Media Conversion Unit.

NOTE 2 – In actual implementations, a Control Processor is typically realized by one or more general purpose CPU(s) and a Media Processor may be for instance a Digital Signal Processing (DSP) device, or a DSP channel in case of a high-capacity DSP device.

NOTE 3 – The qualitative traffic model is applicable for small- and high-capacity Media Gateways. The internal organization of the Media Conversion Units in the MG is outside the scope of this Supplement. There are three main architectural approaches, primarily for H.248 Media Gateways intended for access or core network deployment:

1) circuit interface dedicated MCUs;

2) packet interface dedicated MCUs; or

3) interface independent MCU clusters ("resource pool").

II.3.2.1 Context Processor (CP) and Media Processor (MP) – Service times

The traffic model implies that a MCU is allocated to a H.248 Context for the whole Context lifetime. Thus, the MCU service time $h_{MCU,Context}$ and the MP service time $h_{MP,Context}$ are equal to the Context holding time C_OHT, as shown by Equation II-10.

Mean MCU/MP service time per basic H.248 Context

$$h_{MP,Context} = h_{MCU,Context} = C_OHT \quad [s] \qquad \text{(II-10)}$$

The main relationship between the corresponding service times in control and data paths of a H.248 MG system is:

Ratio between CP and MP service times

$$h_{CP,Context} << h_{MP,Context} \tag{II-11}$$

II.3.2.2 Context Processor (CP) and Media Processor (MP) – Capacity ratio

From Equation II-10, the ideal MCU capacity $\mu_{MCU,Context,max}$ is:

Media Conversion Unit – Ideal service rate $\mu_{MCU,Context,max}$

$$\mu_{MCU,Context,\max} = \frac{1}{C_O HT} \quad [s^{-1}] \tag{II-12}$$

The complete MP Context processing capacity $\mu_{MP,Context,max}$ is given by Equation II-13.

Media Processor – Ideal service rate $\mu_{MP,Context,max}$

$$\mu_{MP,Context,\max} = K \cdot \mu_{MCU,Context,\max} = \frac{K}{C_O HT} \quad [s^{-1}] \tag{II-13}$$

II.3.3 CSN circuit load versus Context holding time

One Media Conversion Unit is needed to serve a single circuit-switched interface. In case of a call, a MCU is allocated to the corresponding CSN interface.[7] A (concentrated or multiplexed) CSN interface is engineered for a mean capacity $A_{CSN,IF,Engineered}$ (also known as link load or concentration factor):

CSN Interface – Engineered load $A_{CSN,IF,Engineered}$

$$A_{CSN,IF,Engineered} = 1 - 0.x \quad [\text{Erl}] \tag{II-14}$$

NOTE – Typical values for $A_{CSN,IF,Engineered}$ are in the range of 0.4 … 0.9 Erlang.

II.3.4 CSN circuit load versus Context control load

The performance between H.248 MG control path and data path has to be appropriately balanced. The underlying design rule is that the system bottleneck may be chiefly the Media Processor. This means that the Context Processor should still have processing resources even when the Media Processor is fully occupied. This engineering concept has a feedback on the H.248 control load.

Based on Equations II-13 and II-14, the meaningful maximum rate of Context Attempts per second $\lambda_{CoAPS,Engineered}$ can be derived as shown in Equation II-15.

Context Processor – Context Attempts per second $\lambda_{CoAPS,Engineered}$

$$\lambda_{CoAPS,Engineered} = A_{CSN,IF,Engineered} \cdot \frac{K}{C_O HT} \quad [s^{-1}] \tag{II-15}$$

The resulting Context Control Processor load $A_{CP,Engineered}$ is (see also Equation 10):

Context Processor – Engineered load $A_{CP,Engineered}$

$$A_{CP,Engineered} = \lambda_{CoAPS,Engineered} \cdot h_{CP,Context} \quad [\text{Erl}] \tag{II-16}$$

---

[7] Circuit-Switched Network (CSN) interface types: analog line, analog trunk, digital line (= ISDN BRI), or digital trunk. H.248 Termination type ALN is intended for analog CSN interfaces, and type TDM is used for digital CSN interfaces.

Equation II-17 gives the corresponding Media Processor load $A_{MP,Engineered}$ (based on Equation II-10):

Media Processor – Engineered load $A_{MP,Engineered}$

$$A_{MP,Engineered} = \lambda_{CoAPS,Engineered} \cdot h_{MP,Context} = \lambda_{CoAPS,Engineered} \cdot C_O HT \quad [\text{Erl}] \qquad \text{(II-17)}$$

In case of a load-balancing mechanism for MCU resources within the MP, the resulting mean Media Conversion Unit load $A_{MCU,Engineered}$ will correspond to:

Media Conversion Unit – Engineered load $A_{MCU,Engineered}$

$$A_{MCU,Engineered} = \frac{A_{MP,Engineered}}{K} \quad [\text{Erl}] \qquad \text{(II-18)}$$

II.3.5 Context Processor performance versus media processor farm size

The Media Processor consists of K Media Conversion Units. The factor K is referred to as 'farm size' parameter.

The theoretical maximum capacities in control and data path are:

- Context Processor: $A_{CP,max} = 1$ Erl (for the single server model)
- Media Processor: $A_{MP,max} = K$ Erl (for the K-server model)

The engineered CSN link load $A_{CSN,IF,Engineered}$ typically results from network planning, for instance, engineering a link for certain grade of service parameters (like blocking probability). For specific MP architectures, the farm size factor may be reduced by benefiting from economy of scales effect.

II.3.6 Calculation examples

This clause shows some examples of interrelationships among User plane capacity, MG data path size, and MG control performance.

II.3.6.1 MG size variation: $\phi_{CoCPS} = f(K)$

The size of MGs may vary from small to high capacity systems. The size factor affects the dimensioning of the data and control paths. Farm size factor K is the prime data path parameter for C2X MG types.

How the required control performance of the H.248 Context Processor depends on the MG size is defined by Equation II-15. If we combine this relation with the fact that every Context Attempt must be completed, this leads to following functional behaviour $\phi_{CoCPS} = f(K)$.

Context Processor Performance as a function of K

$$\phi_{CoCPS,Engineered}(K) = \frac{A_{CSN,IF,Engineered}}{C_O HT} \cdot K \quad [\text{s}^{-1}] \qquad \text{(II-19)}$$

The control performance is linearly related with the CSN interface capacity, under the assumption that the concentration factor $A_{CSN,IF,Engineered}$ and the Context holding time $C_O HT$ are constant.

II.3.6.2 Link load variation: $\phi_{CoCPS} = f(A_{CSN,IF})$

Equation II-20 also provides the dependency of engineered concentration level at MG circuit interfaces:

Context Processor performance as a function of $A_{CSN,IF}$

$$\phi_{CoCPS,Engineered}(A_{CSN,IF}) = \frac{K}{C_O HT} \cdot A_{CSN,IF,Engineered} \quad [\text{s}^{-1}] \qquad \text{(II-20)}$$

The control performance is linearly related with the CSN interface concentration level, under the assumption that the MP farm size K and the Context holding time C_OHT are constant.

II.3.6.3 Context Holding Time variation: $\phi_{CoCPS} = f(C_OHT)$

The probability distributions functions for Context holding times are dependent of many parameters. Equation II-21 provides also the principle dependency of the control performance from data path resource holding times:

Context Processor performance as a function of C_OHT

$$\phi_{CoCPS,Engineered}(C_OHT) = K \cdot A_{CSN,IF,Engineered} \cdot \frac{1}{C_OHT} \quad [s^{-1}] \tag{II-21}$$

The control performance is hyperbolically related with the mean Context holding time, under the assumption that the MP farm size K and the concentration factor $A_{CSN,IF,Engineered}$ are constant. This non-linear behaviour is elaborated in II.4.

II.4 Effective throughput versus Context Holding Time: $\phi_{CoCPS} = f(C_OHT)$

The H.248 Context holding times are very service-, market-, and/or operator-specific. Varying the mean holding time impacts the Context Processor performance. The overload control model of II.2 allows the derivation of the main behaviour.

II.4.1 Derivation

The derivation of the functional relationships is based on the framework given in II.2.

II.4.2 Results

The mean Completion rate for H.248 Contexts as a function of the H.248 Context holding time, $\phi_{CoCPS} = f(C_OHT)$, is given by Equation II-22 for the three workload areas of the H.248 Context Control Processor:

Context throughput $\phi_{CoCPS} = f(C_OHT)$; with blind load handling; including static overhead and reserves

$$\phi_{CoCPS} = f(C_OHT) = \begin{cases} \lambda_{CoAPS} = \dfrac{1}{C_OHT} & \text{for } C_OHT \geq \hat{h}_{CoC} \quad \text{Underloaded Server} \\ \dfrac{1-(\rho_{BL}+\rho_{HR})}{(1-\kappa)h_{CoC}} - \dfrac{\kappa}{1-\kappa} \cdot \dfrac{1}{C_OHT} & \text{for } \hat{h}_{RC} \leq C_OHT \leq \hat{h}_{CoC} \quad \text{Overloaded Server} \\ 0 & \text{for } C_OHT < \hat{h}_{CoR} \quad \text{Unstable Server} \end{cases} \tag{II-22}$$

NOTE 1 – The differences between Equations II-21 and II-22 are that Equation II-21 is only valid for an underloaded Context Processor, and derived from the specific control/data path traffic model for Circuit-to-X H.248 Contexts, whereas Equation II-22 is fairly general because it only considers the MG control path. Equation II-22 is even applicable as a model for an MGC-level Context Processor.

The boundary values $\hat{h}_{CoC}$ and $\hat{h}_{CoR}$ are given by Equations II-23 and II-24, respectively:

Limit parameter $\hat{h}_{CoC}$

$$\hat{h}_{CoC} = \frac{1}{1-(\rho_{BL}+\rho_{HR})} h_{CoC} \tag{II-23}$$

Limit parameter $\hat{h}_{CoR}$

$$\hat{h}_{CoR} = \frac{1}{1-(\rho_{BL}+\rho_{HR})} h_{CoR} \tag{II-24}$$

Figure II.6 illustrates the functional behaviour characterized by Equation II-22.

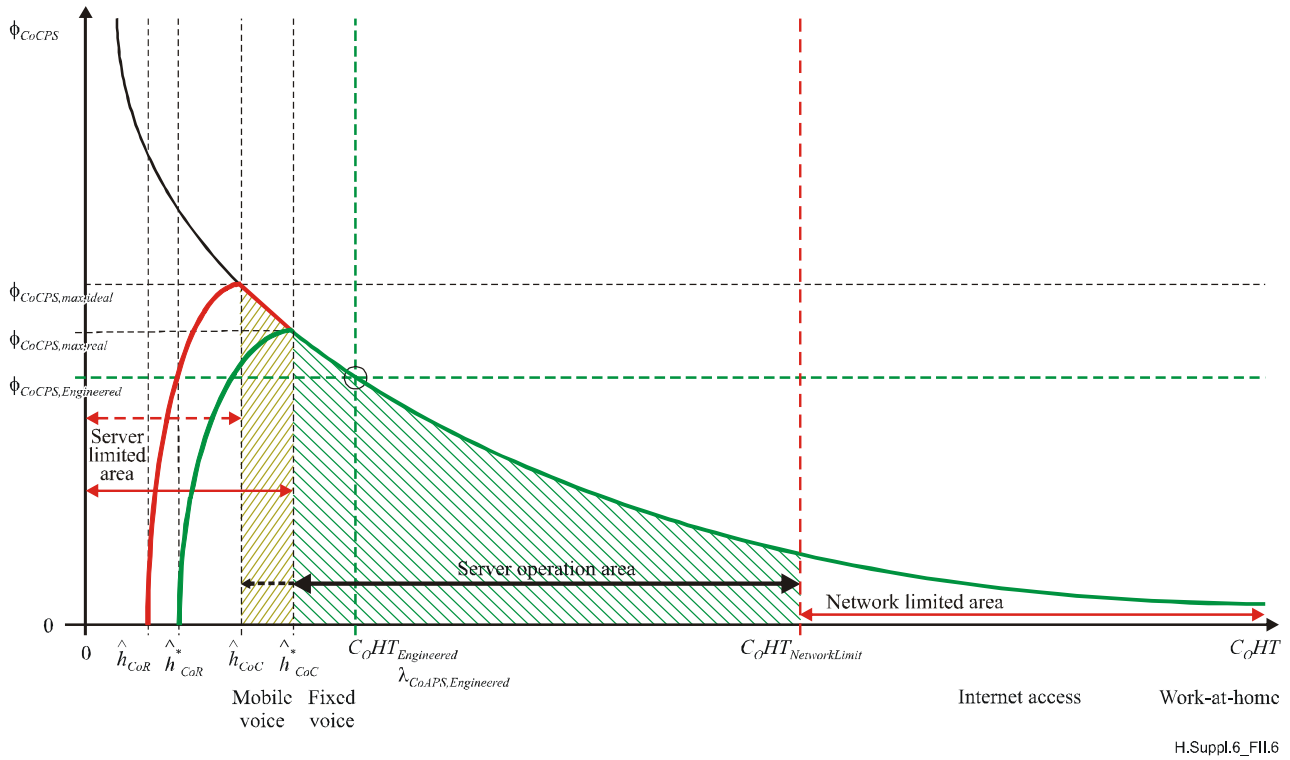

Figure II.6 – Recommended operation area for H.248 Context Processor

NOTE 2 – At the bottom of Figure II.6, some qualitative values for mean Context holding times of several services are pointed out. Typically, $C_OHT_{MobileVoice} < C_OHT_{FixedVoice} < C_OHT_{InternetAccess} < C_OHT_{Work\text{-}at\text{-}Home}$ for the expectation values of the corresponding underlying probability distribution functions.

The system is engineered for the operation point $\{C_OHT_{Engineered} \mid \lambda_{CoAPS,Nominal}\}$, where $\lambda_{CoAPS,Nominal}$ (or $\lambda_{CoAPS,Engineered}$) specifies the nominal load or engineered capacity (in terms of the Context Attempt arrival rate).

II.4.3 Conclusions

Equation II-22 may be interpreted in the following ways:

- Strong non-linear dependency of achievable Context processing capacity versus average Context holding time (C_OHT).
- Range of applicable average C_OHTs is limited by the theoretical maximum system capacity and engineered network capacity.
- Linear relationship assumptions are only applicable for "very small" C_OHT ranges. Linearization should be applied with utmost caution.
- Regarding network engineering, the uncertainties concerning support of wider ranges of average C_OHT values (e.g., due to specific service distribution, call mixes, etc.) have to be supported by broader scalability ranges of Context Processor capacities.
- There is a hyperbolical relationship between effective throughput and holding time in the normal operation mode of the Context Processor ('underload' state).

The useful Context Processor operation area is bounded by the constraints of network and system limitations.

NOTE 3 – More background on network limited area and system limited area is indicated in GR-517-CORE [17]; see Figure 5-3 of GR-517-CORE.

II.5 Overload Control Model for access gateways

II.5.1 Background and applicability statements

The model(s) may be used in following network context:

- PSTN/ISDN Emulation Subsystems (PES);
- Access network side (interfaces with legacy terminals and/or PBXs);
- VoIP NGN (call/session control protocol is, e.g., SIP).

The model(s) may be used in following service (traffic) context:

- consideration of emergency telecommunication services (ETS) besides non-ETS calls;
- focus on calls originating at access side;
- incoming calls from (core) network side (optional).

The model(s) may be used for the following H.248 functions:

- MGC Overload Protection by MG (for PSTN calls only);
- MG Overload Control according to ITU-T Rec. H.248.10; and/or
- MG Overload Control according to ITU-T Rec. H.248.11.

The access gateways comprise a pair of H.248 master-slave entities:

- H.248 MGC (e.g., AGCF); and
- H.248 MG (e.g., Residential MG, Access MG).

The following clauses provide overall models for access gateways. Each model may be decomposed in case of dedicated performance investigations.

II.5.2 PSTN-only model

Figure II.7 shows the model based on the network architecture, which is related to a functional architecture. The H.248 Access Media Gateway (AMG) interfaces Analog Lines (ALN) with the IP network. The H.248 ALN Termination is used for bearer traffic and call control traffic. Call control protocols are summarized by the term "Analog Line Signalling" (ALS). ALS is generally "pre-processed" by the H.248 MG and forwarded to the H.248 MGC (e.g., by E.9/H.248.1). The MGC is the primary instance for call control activities.

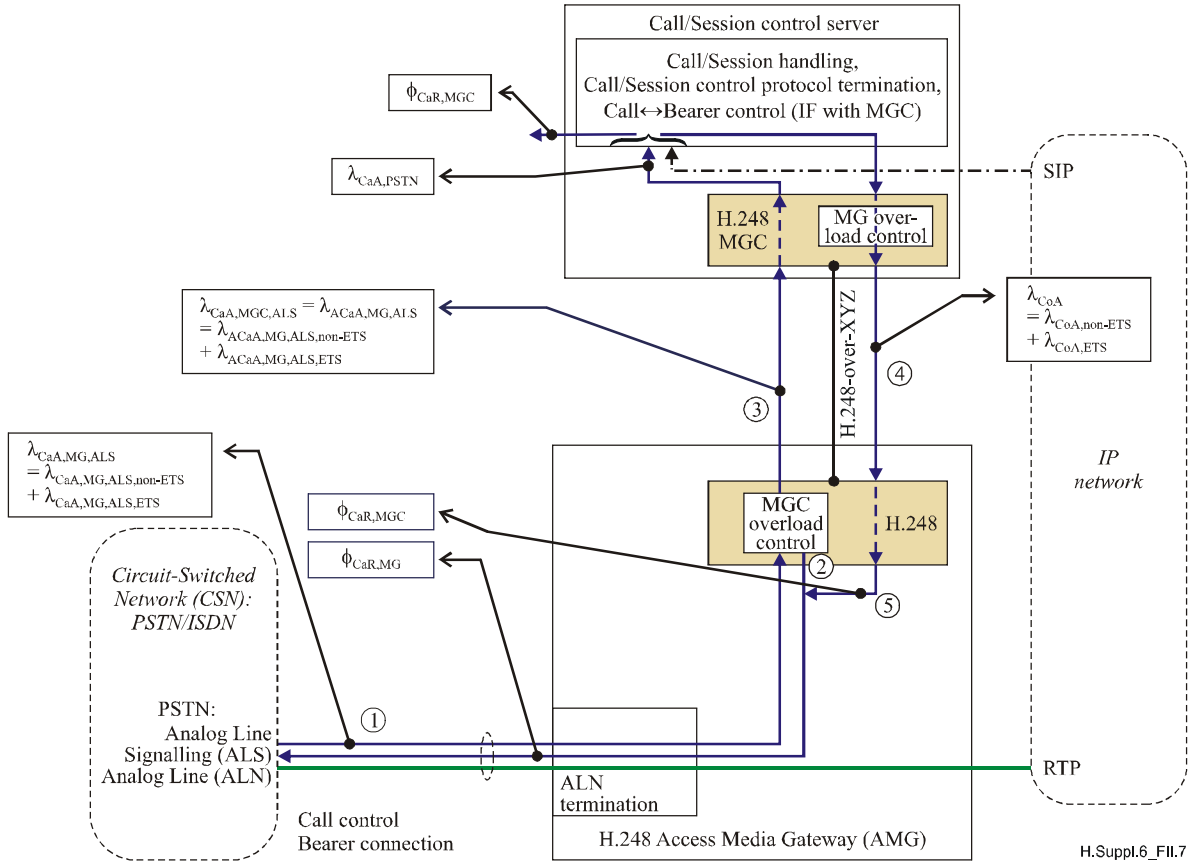

Figure II.7 – H.248 Access gateways – PSTN-only model

The various traffic variables in Figure II.7 are introduced by following the flow of a call originating in the PSTN access network. The first four major stages are:

1) Call arrival rate $\lambda_{CaA,MG,ALS}$ represents all call attempts on the MG level. This is typically related to the "supervisory signalling" event "off-hook".

   Emergency calls are denoted by $\lambda_{CaA,MG,ALS,ETS}$; other calls are summarized by rate $\lambda_{CaA,MG,ALS,non-ETS}$.

2) Call rejection rate $\phi_{CaR,MG}$ represents all rejected call attempts by the MG itself. This MG capability is linked to a specific "MGC overload protection scheme" (e.g., Ref: to be included). "Rejection" may be related to a "congestion tone", "missing dial tone", etc.

3) Call arrival rate $\lambda_{CaA,MGC,ALS}$ represents all call attempts on the MGC level. This rate is synonym to the MG-accepted call attempt rate $\lambda_{ACaA,MG,ALS}$. Again there is a distinction between emergency and non-emergency calls ($\lambda_{CaA,MGC,ALS,ETS}$ and $\lambda_{CaA,MGC,ALS,non-ETS}$).

4) Context arrival rate λ_{CoA} represents all H.248 Context Attempts from MGC to MG. This rate relates to all "accepted call attempts" by the call control on MGC (or higher) level. An H.248 Context may be attributed with regard to emergency services. This is reflected by the two sub-rates $\lambda_{CoA,ETS}$ and $\lambda_{CoA,non-ETS}$.

II.5.3 PSTN/ISDN model

The previous PSTN model is solely considering analog line interfaces with the MG. The PSTN/ISDN model (Figure II.8) additionally covers ISDN interfaces like BRI (or PRI). These ISDN interfaces are User-Network Interfaces (UNI) with call control signalling according to DSS1. The term "xSS1" indicates that other "DSS1-related" call control protocols are also in scope (e.g., PSS1, DPNSS1, DASS1, QSIG, etc.).

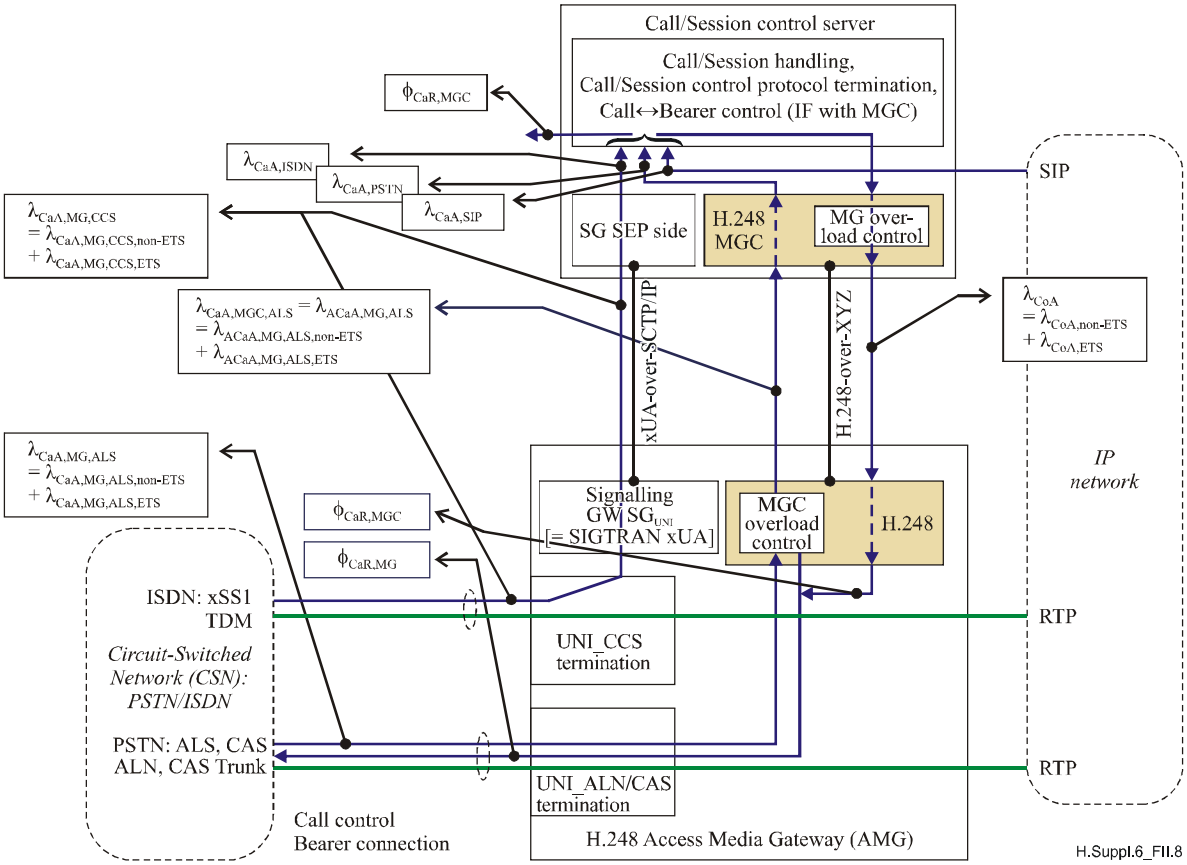

Figure II.8 – H.248 access gateways – PSTN/ISDN model

The call control protocol "xSS1" belongs to Common Channel Signalling (CCS). Any kind of CCS-based call control (where CCS is of type FAS[8]) is handled by H.248 AMG via embedded Signalling Gateways (SGs). The considered SG types are based on IETF SIGTRAN solutions (indicated by the term "xUA" for IUA, or DUA).

The SG and H.248 MG functions are disjoint in the control plane. The call arrival rate $\lambda_{CaA,MG,CCS}$ on "SG/MG" level is therefore identical to the MGC level.

II.6 Overload Control Model for ITU-T Rec. H.248.11

II.6.1 Background

ITU-T Rec. H.248.11 defines a feedback-based, closed control. The control loop spans the two H.248 MGC and MG entities; thus, it is equivalent to a so-called external overload control. Therefore, the model basically comprises a tandem of a single MGC-MG pair.

ITU-T Rec. H.248.11 is designed for virtual MG (VMG) support. An extension of the basic model may be a configuration with multiple MGC-VMG pairs (see II.6.3).

II.6.2 Basic H.248.11 model for a single MGC-MG pair

Figure II.9 shows the basic model with H.248 interface and the overlaid control loop. Any control may be decomposed in characteristic components. The proposed model distinguishes four components (A, D, R, U), according to the NGN overload control architecture as defined by ETSI TISPAN TR 182 015.

8 The SG may be MG-external in case of type "Non-FAS" (NFAS).

The variables of the H.248.11-based control, which are highlighted in Figure II.9, are:

- Event notification rate ε, based on the notification of H.248.11 Event ocp/mg_overload; and
- TargetMG_OverloadRate δ (as defined in 8.2.3/H.248.11).

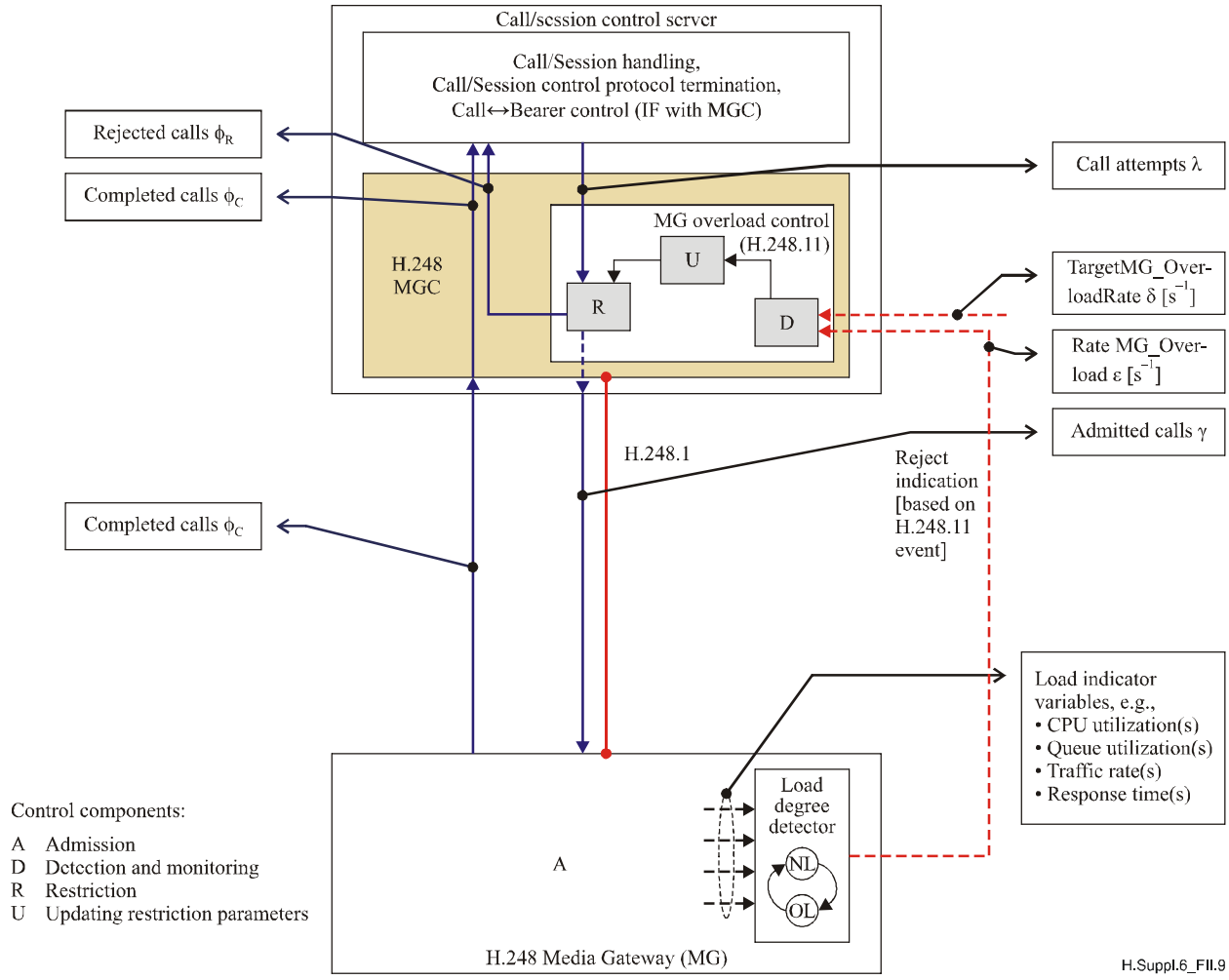

Figure II.9 – H.248 gateway – Basic model for H.248.11

The various traffic variables in Figure II.9 are introduced by following the flow of a new call attempt. The first four major stages are:

1) Call arrival rate or call attempt rate λ_{CaA} (denoted λ in Figure II.9) represents all call attempts on the MGC level. The call originates in the served user instance of the MGC (e.g., call/session handling block in Figure II.9). This instance may be abstracted by a traffic source model.

2) Call rejection rate ϕ_{CaR} (denoted ϕ_R in Figure II.9) represents all rejected call attempts by the MGC, based on H.248.11 load regulation.

3) Admitted call rate γ corresponds to the Context arrival rate λ_{CoA}

   NOTE – γ is used here according to the H.248.11 terminology.

4) Context and call completion rate ϕ_{CoC} and ϕ_{CaC} are identical in this basic model, thus abbreviated as ϕ_C in Figure II.9.

The restriction component is a load regulator based on a leaky bucket type (see 3.5/H.248.11). The leaky bucket itself is not highlighted in Figure II.9, but taken as an inherent part of the model.

II.6.3 Model with virtual MG support

For further study.

II.6.4 Additional modelling of loss in MG

The MG is lossless in the basic model discussed II.6.2. The loss-free property is reflected in the model by the equality of the stationary values of γ and ϕ_C. The model may be extended to consider additional rejection of Context Attempts (e.g., by variable Context rejection rate ϕ_{CoR}).

Appendix III

Examples of control processing capacity computations

For further study.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	**Audiovisual and multimedia systems**
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2006