

Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS/NGN Performance Benchmark Part 1: Core Concepts



Reference

DTS/TISPAN-06024-1-NGN

Keywords

IMS, performance, service

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

Individual copies of the present document can be downloaded from:

<http://www.etsi.org>

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the PDF version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at

<http://portal.etsi.org/tb/status/status.asp>

If you find errors in the present document, please send your comment to one of the following services:

http://portal.etsi.org/chaicor/ETSI_support.asp

Copyright Notification

No part may be reproduced except as authorized by written permission.
The copyright and the foregoing restriction extend to reproduction in all media.

© European Telecommunications Standards Institute 2007.
All rights reserved.

DECTTM, **PLUGTESTS**TM and **UMTS**TM are Trade Marks of ETSI registered for the benefit of its Members.
TIPHONTM and the **TIPHON logo** are Trade Marks currently being registered by ETSI for the benefit of its Members.
3GPPTM is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners.

Contents

Intellectual Property Rights	4
Foreword.....	4
Introduction	4
1 Scope	7
2 References	7
2.1 Normative references	7
2.2 Informtaive references.....	8
3 Definitions and abbreviations.....	8
3.1 Definitions	8
3.2 Abbreviations	10
4 Benchmark information model.....	11
4.1 Use-case	11
4.1.1 Metrics	12
4.1.2 Use-case outcomes.....	12
4.1.3 Scenarios and scenario attempts	12
4.1.4 Design Objective (DO)	12
4.1.5 Duration distribution.....	13
4.1.6 Scenario	13
4.2 Benchmark test	13
4.2.1 Traffic set.....	13
4.2.2 Background load.....	14
4.2.3 Arrival distribution	14
4.2.4 Traffic-time profile	14
4.2.5 Test parameters	14
4.3 Benchmark report.....	15
4.3.1 Disclaimer.....	15
4.3.2 SUT configuration and parameters	15
4.3.3 Test system configuration and parameters.....	16
4.3.4 Report of metrics	16
4.3.5 Observations and interpretations.....	16
4.3.6 Exceptions.....	16
4.3.7 Relationship between system load and number of subscribers	16
5 System Under Test (SUT)	17
5.1 General SUT requirements	18
5.2 Creation of subscriber base data	18
6 Test system.....	19
7 Benchmark metrics.....	19
Annex A (informative): Illustrative example of benchmark report.....	22
A.1 Benchmark	22
A.2 Benchmark results	22
A.2.1 Registration traffic (time series) (overall)	23
A.2.2 Average duration of completed registrations (time series) (overall)	24
A.2.3 Histogram of delays between consecutive REGISTER requests (Histogram) (overall).....	24
Annex B (informative): TTCN-3 based test system sample.....	26
Annex C (informative): Bibliography.....	27
History	28

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (<http://webapp.etsi.org/IPR/home.asp>).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN).

The present document is part 1 of a multi-part deliverable covering the IMS/NGN Performance Benchmark, as identified below:

Part 1: "Core Concepts";

Part 2: "Subsystem Configurations and Benchmarks".

Part 3: "Traffic Sets and Traffic Profiles".

Introduction

A major change is occurring in the telecommunications domain. Telecom service providers (SPs) are quickly evolving their networks from legacy technologies to what might be termed "fourth generation" or "3G Beyond" technologies. The principal features of this technological change are:

- Evolution of "traditional" wireline telecom standards to Voice over IP (VoIP) standards, pre-eminently the cluster of protocols surrounding SIP.
- Evolution of GSM and CDMA networks to 3GPP/3GPP2 standards, such as UMTS.
- Introduction of wireless LAN (WLAN) standards, pre-eminently IEEE 802.16 [5], for both data and voice communications.
- Fixed-mobile convergence through the various access technologies standardized by TISPAN.
- Potential and actual competition among standards, causing in reaction an attempt at service convergence.

The current convergence point of these trends is at a set of technologies termed the IP Multimedia Subsystem (IMS). IMS supports a rich set of services available to end users on either wireless or wired user equipment (UE), provided via a uniform interface by the cooperation between a subscriber's home service provider (SP) and the visited service providers. Services are provided via an "overlay" technique over multiple service provider networks.

Telecom Equipment Manufacturers (TEMs) all along the architectural hierarchy, assuming that IMS represents a growth market, are attempting to develop not merely products for IMS networks, but architectures as well. The quest for new architectures represents a view that current processor, network, and server architectures are not sufficient to support wide IMS deployment. Examples of such work include:

- Advanced Telecom Architecture (ATCA), a set of existing and emerging standards for the physical packaging of bladed servers and communications fabrics.
- Fabric standards, such as Infiniband, PCI Express, RapidIO and Gigabit Ethernet (e.g. IEEE 802.1 and IEEE 802.3).

- Technologies for offloading packet processing, such as network processors and other multicore processors.
- Real-time and carrier-grade operating systems, such as the carrier-grade Linux releases of Montavista and Red Hat.
- Middleware architected for high availability, such as defined by the Object Management Group (OMG) and the Service Availability Forum (SAF).
- Application development environments and interfaces, such as defined by the Parlay Group and the Java Community Process (e.g. JAIN).

The number of technological variables is so large that some reasonable ground rules for defining an architecture need to be put in place. SPs require guidance for making decisions among suppliers, and suppliers all along the architectural hierarchy need guidance to develop the right products.

- The present document provides a general introduction to the environment in which the benchmark exists.
- TS 186 008-2 [3] documents the subsystem configurations, use-cases and design objectives corresponding to them.
- TS 186 008-3 [4] documents the benchmark tests through definitions of traffic sets and traffic-time profiles.

Motivating Example

Figure 1 depicts in a single diagram the major features of the IMS performance benchmark, and will serve as a motivating example. The System Under Test (SUT) is presented with a workload by the Test System. The workload consists of the traffic generated by a large number of individual simulated User Endpoints (UEs), each performing an individual scenario. The collection of scenarios selected for a benchmark test is a traffic set.

The rate at which scenarios from the traffic set are attempted in the benchmark test is governed by a traffic-time profile defined for the benchmark. A traffic-time profile is designed so that the rate of scenario attempts remains constant for sufficient time to collect a statistically significant data set. For example, in the figure, the time between 0 minute and 10 minutes shows the traffic rate ramping up to approximately 120 session attempts per second (SAPS), after which it remains in a steady-state phase for approximately 30 minutes, at an average value of 120 SAPS.

During the test, each scenario attempt (a generalization of "call attempt", accounting for the fact that a scenario may be a registration scenario or an instant messaging scenario) may succeed, may fail, or may succeed but exceed the time threshold determined to be the "design objective" for the scenario. During the steady state phases of a traffic-time profile, the percentage of failures or timeouts, collectively called "inadequately handled scenario attempts", is measured. If it exceeds a certain frequency, which may also be interpreted as a probability of inadequately handled scenario attempt, the SUT has reached its Design Objective Capacity (DOC), the primary comparison metric reported by the benchmark. In figure 1, the DOC is reached at 130 SAPS.

When the traffic-time profile changes to a new average value, the percentage of inadequately handled scenario attempts will in general exhibit a transient spike. The average successful scenario rate and the percentage of inadequately handled scenario attempts is averaged over the steady-state phase excluding the transient.

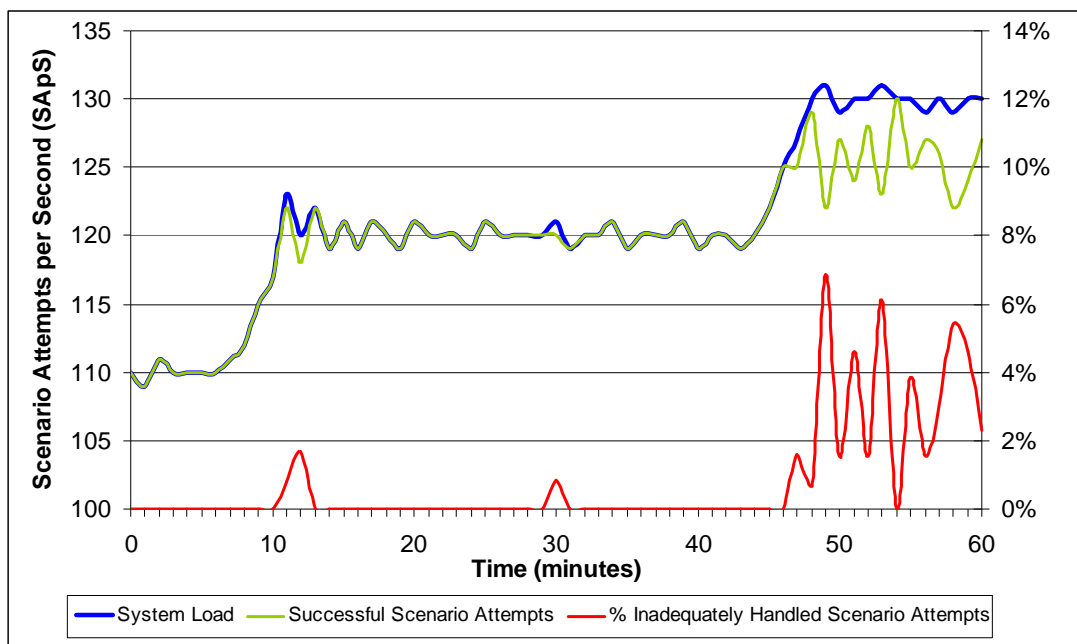


Figure 1: Motivating example of an IMS benchmark

1 Scope

The present document is for an initial, revision 1.0, release of an IMS/TISpan NGN performance benchmark. The metrics measured and reported are for performance of this subsystem under a communications application load.

The benchmark is defined for the IMS network as a whole, as well as for several subsystems of an IMS network. The benchmark is designed so that nodes composing a subsystem can also be benchmarked alone.

This multi-part deliverable consists of three parts document is in three parts. The present document contains overall benchmark descriptions, architectures, processes, and information models that are common to all specific benchmarking scenarios.

TS 186 008-2 [3] contains the IMS and ETSI TISpan SUT subsystem configurations, the specific benchmarking use-cases and scenarios, along with scenario specific metrics and design objectives. It also defines the SUT configuration parameters. This part also contains any required extensions to the overall descriptions present in the present document, if necessary for the specific scenario.

TS 186 008-3 [4] defines an initial benchmark test through the specification of a traffic set, traffic-time profile and benchmark test procedure.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific.

- For a specific reference, subsequent revisions do not apply.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
 - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
 - for informative references.

Referenced documents which are not found to be publicly available in the expected location might be found at <http://docbox.etsi.org/Reference>.

For online referenced documents, information sufficient to identify and locate the source shall be provided. Preferably, the primary source of the referenced document should be cited, in order to ensure traceability. Furthermore, the reference should, as far as possible, remain valid for the expected life of the document. The reference shall include the method of access to the referenced document and the full network address, with the same punctuation and use of upper case and lower case letters.

NOTE: While any hyperlinks included in this clause were valid at the time of publication ETSI cannot guarantee their long term validity.

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

- [1] ETSI ES 282 007 (V1.1.1): "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISpan); IP Multimedia Subsystem (IMS); Functional architecture".
- [2] ITU-T Series H, Supplement 6 (April 2006): "Control Load Quantum for Decomposed Gateways".

- [3] ETSI TS 186 008-2: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS/NGN Performance Benchmark Part 2: Subsystem Configurations and Benchmarks".
- [4] ETSI TS 186 008-3: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS/NGN Performance Benchmark Part 3: Traffic Sets and Traffic Profiles".
- [5] IEEE 802.16: "A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access".

2.2 Informative references

- [6] ETSI TS 129 163: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Interworking between the IP Multimedia (IM) Core Network (CN) subsystem and Circuit Switched (CS) networks (3GPP TS 29.163 version 7.7.0 Release 7)".

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply:

arrival distribution: function describing the probability distribution of the interarrival time of events (e.g. messages)

background load: workload applied to an SUT during a benchmark test, for the purpose of consuming SUT resources during a benchmark test and changing the traffic intensity at which the capacity of the SUT is reached

benchmark log: data file containing measurements of SUT performance collected during the execution of a test procedure

benchmark report: documented generated at the conclusion of a test procedure containing the metrics measured during the execution of the test and/or computed from the data collected in the benchmark log

benchmark test: procedure by which a test system interacts with a System Under Test to measure its behavior and produce a benchmark report

configuration: specification of a subset of IMS architectural elements and metrics for which collection of benchmark tests can be defined

design objective: probabilistic model of delay and failure requirements for an SUT, associated with a use-case, specified by threshold values and probabilities for delay and scenario failure.

Design Objective Capacity (DOC): largest load an SUT can sustain while not exceeding design objectives defined for a use-case

DOC overload: condition, or part of a load profile, in which the system load exceeds the DOC

DOC underload: condition, or part of a load profile, in which the system load does not exceed the DOC

duration distribution: function (e.g. Poisson) describing the probability distribution of the duration of an event (e.g. a scenario)

inadequately handled scenario attempt: scenario attempt which either fails or which exceeds the threshold values defined for the use case of which the scenario attempt is an instantiation

maximum capacity: smallest scenario arrival rate at which the successful scenario rate cannot be increased

metric: performance measurement of an SUT reported in a benchmark report

parameter: attribute of an SUT, test system, system load, or traffic set whose value is set externally and prior to a benchmark test, and whose value affects the behavior of the benchmark test

percent registered subscribers: parameter specifying the percent of subscribers with records in the HSS/UPSF that are active during the benchmark test

percent roaming subscribers: parameter specifying the percent of active subscribers who are roaming

preamble: phase at the beginning of a test procedure during which initialization of the test system and system under test is performed

protocol diagram: diagram depicting a collection of architectural elements as vertical lines, and protocol interactions between the elements as directed lines between architectural elements, where the vertical order in which the directed lines appear indicate time sequence

Protocol Implementation eXtra Information for Testing (PIXIT): statement made by a supplier or implementor of an IUT (protocol) which contains or references all of the information related to the IUT

recovery capacity: when a traffic-time profile describes scenario arrival rate starting at a value greater than maximum capacity and monotonically decreasing, the maximum value at which design objectives for the use-case in effect are no longer exceeded

SAPS increase amount: increment by which the average SAPS changes between steps of a profile

NOTE: Equivalent to system load increase amount.

scenario: specific path through a use-case, including the sequence of messages exchanged by all agents, and (when meaningful) a scenario duration distribution (e.g. the duration of a scenario)

NOTE: An example of a scenario is "simple call - succeeded".

scenario attempt: event of a scenario being initiated by the test system and handled by the SUT

scenario attempts per second: average number of scenarios that are instantiated by the test system per second

scenario arrival distribution: probability distribution that governs the arrival times of scenarios during a test phase

scenario duration distribution: probability distribution that governs the duration of an individual scenario

scenario percent of system load: relative frequency of an individual scenario within a system load

simultaneous scenarios: number of scenarios that the test system may allow a single UE to perform simultaneously

step number: for a profile consisting of a sequence of steps, the number of steps

step time: length of time, in a profile consisting of a sequence of steps, at which the average scenario arrival rate remains at the same value

step transient time: parameter representing the time interval, measured from the beginning of a step, for which counts of scenario attempts and inadequately handled scenario attempts are not kept

stir time: parameter representing a period of time in the preamble of a benchmark test in which a system load is run in order to allow initial transient conditions attenuate to an insignificant level

subscriber base: information elements that describe simulated users

system load: stream of protocol interactions presented to the SUT by the test system

system load increase amount: increment by which the average SAPS changes between steps of a profile

NOTE: Equivalent to SAPS increase amount.

System under Test (SUT): collection of hardware and software whose performance is measured by the benchmark test

test parameters: parameters whose values determine the behavior of a benchmark test

test procedure: specification of the steps to be performed by a benchmark test

test scenario: specific path through a use-case, whose implementation by a test system creates a system load

test system: collection of hardware and software which presents a system load to a system under test and collects data on the system under test's performance, from which metrics can be computed

total provisioned subscribers: number of simulated subscribers with records in the HSS/UPSF

traffic-time profile: evolution of the average scenario arrival rate over time

NOTE: It is specified by a scenario arrival distribution and a function of average scenario arrival rate as a function of time.

traffic set: mixture of scenarios whose proportional contributions to traffic are fixed

use-case: specification of a type of interaction between a test system and a system under test, corresponding to a mode of end-user behavior

NOTE: A use-case is a collection of scenarios.

EXAMPLE: An example of a use-case is "Registration/De-registration", which may contain scenarios "Successful Initial Registration without Synchronization", "Re-Registration - User Roaming", and others.

user behavioral model: model that defines the number and rate at which an individual user of an IMS system makes scenario attempts

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

%IHS	Percent Inadequately Handled Scenarios
3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project-2
APS	Attempts Per Second
AS	Application Server
ATCA	Advanced Telecom Architecture
ATCA	Advanced Telecom Computing Architecture
CDMA	Code Division Multiple Access
CSCF	Call Session Control Function
DO	Design Objective
DOC	Design Objective Capacity
DOC	Design Objective Capacity
GSM	Global System for Mobile communication
HSS	Home Subscriber Server
IHSA	Inadequately Handled Scenario Attempt
IHSA	Inadequately Handled Scenario Attempt
IMS	IP Multimedia Subsystem
JAIN	Java API for Integrated Networks
LAN	Local Area Network
NGN	Next Generation Networks
PCI	Peripheral Component Interconnect
P-CSCF	Proxy-CSCF
PIXIT	Protocol Implementation eXtra Information for Testing
PSTN	Public Switched Telecommunications Network
SAF	Service Availability Forum
SAPS	Scenario Attempts Per Second
SAPS	Session Attempts Per Second
SIMS	SIMultaneous Sessions (per user)
SUT	System Under Test
TEM	Telecom Equipment Manufacturer
TISPAN	Telecommunications and Internet convergence Services and Advanced Networking
UE	User Equipment
UPSF	User Profile Server Function

4 Benchmark information model

In this clause, "benchmark information model" refers to the structure of the information elements that define the benchmark. This information model is depicted in figure 2.

The benchmark information model takes the approach that any workload presented to an SUT starts with the behavior of an individual user. When a user interacts with the SUT, he does so with a particular goal (e.g. to make a call or to send a message). The SUT may provide a variety of ways to accomplish this goal, but the high-frequency actions are relatively limited in number, and can be captured by a manageable set of scripts. Individual users may differ in the relative speed with which the actions are performed, but this behavior can be described via a probability model.

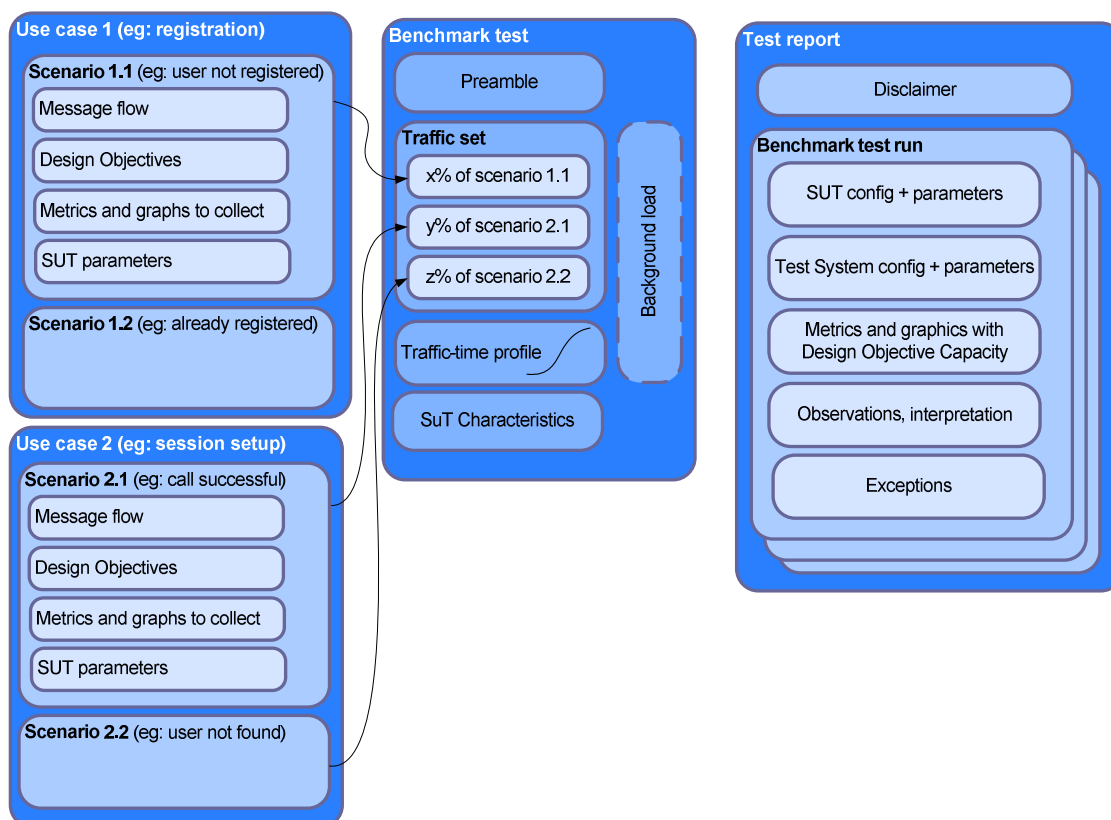


Figure 2: IMS benchmark information model

The information model consists of three primary elements: use-cases, which describe the behavior of an individual user, and which in turn define scenarios; benchmark tests, which generate a workload by aggregating the behavior of individual scenarios in a controlled manner, and collect logfiles of measurements during the test; and benchmark test reports, which report metrics interpreted from the benchmark test log files.

The remainder of this clause will describe the use-case, benchmark test, and benchmark report in more detail.

4.1 Use-case

The top level of the individual behavioral model is the *use-case*. A use-case describes the goal that a user has in interacting with a system, the various actors (e.g. other users, network elements) that participate in the use-case, the basic course of events that are carried out by the user and the SUT, the design objective of the use-case, the possible outcomes that apply to the use-case, and the metrics to be collected. The goal and actors of a use-case are documented in narrative text and diagrams; the other elements are complex information elements, which are described in their respective clauses.

4.1.1 Metrics

The metrics of a use-case describe the measurements collected from the execution of a scenario attempt. Typical metrics include response times and message rates. If a scenario is selected for execution in a benchmark test, its metrics are collected. See clause 7 for more detail.

4.1.2 Use-case outcomes

A use-case outcome is a set of possible outcomes of the scenarios of a use-case. An outcome may be simply "correct", it may reflect an error or failure condition; or it may reflect a correct behavior that took an excessive amount of time to occur. An instance of a scenario that experiences an error, failure, or timeout outcome is referred to as an inadequately handled scenario attempt.

4.1.3 Scenarios and scenario attempts

A scenario is a trace of a path through a use-case. It is analogous to "call attempt" (see also discussion in clause 3.1 in [2]), but applies to all interactions within an IMS network, such as registrations, text messages, and application interactions.

A scenario attempt is the event of a scenario being executed in an IMS network. A scenario may succeed, fail, or succeed functionally but take longer than the time thresholds associated with its use-case allows. In the latter two cases, the scenario attempt is deemed an "Inadequately Handled Scenario Attempt" (IHSA).

The term "scenario attempt" and "scenario attempts per second" are used in this standard in place of "call attempt" and "call attempts per second" because IMS is a transaction-oriented system with transactions of a variety of types (e.g. calls, registrations, deregistrations and text messages, etc.). Traffic sets, and indeed the real world, do not operate according to only one transaction type, so the more generalized term is necessary. It would be incorrect and misleading to attempt to report the capacity of a system in "call attempts per second", "registration attempts per second", etc., for system loads that were other than purely call attempts, registration attempts, etc.

The term "scenario attempt" is inline with the terms "session attempt" and "context attempt" according [2], which is applicable for decomposed gateway components within IMS/NGN. See for instance the gateway functions related to Mp and Mn in figure 4.

Scenario attempts could be further categorized into call-dependent (e.g. conversational services, streaming services) and call-independent (e.g. registration, and roaming, etc.) scenario attempts.

NOTE: The notion of call-dependent/call-independent is also used e.g. in TS 129 163 [6].

This categorization is only meaningful for network elements that are able to differentiate both scenario categories (e.g. P-CSCF).

Similarly, when measuring timeout thresholds, this standard uses specific terms such as "TRT-REG" (Total RoundTrip Time-Registration) measured between precisely defined reference points, rather than the catchall term "setup delay", in order to allow measurement to be carried out in an unambiguous manner.

4.1.4 Design Objective (DO)

The Design Objective (DO) describes the acceptable rate of inadequately handled scenario attempts for a use-case. When a benchmark test is executed, scenarios from various use-cases are selected and executed, thereby becoming "scenario attempts". Any of the scenarios that are deemed to be inadequately handled add to the count of inadequately handled scenario attempts (IHSAs). If the frequency of IHSAs exceeds the design objective, then the Design Objective Capacity (DOC) has been exceeded.

DO, IHSA, and DOC are necessary concepts to define an end-to-end benchmark, such as is described in this standard. However, a fully-agreed model of thresholds for IHSAs, and a DO model will not emerge for some time. For this reason, benchmark test definitions in terms of IHSA and DO models based on current knowledge are maintained separately in TS 186 008-3 [4]. As industry and standards organizations gain experience in IMS deployments, these models will be refined.

4.1.5 Duration distribution

The duration distribution is a probability model describing the durations of the interactions between the actors of a use-case. The model is applied when a scenario is performed by a benchmark test, but because it is associated with the actors of a use-case, particularly to the user, it is common among all scenarios of a use-case.

4.1.6 Scenario

A scenario describes a single interaction sequence among the actors of a use-case. It is documented by a set of preconditions on its actors (typically specified by parameter values), a protocol diagram describing the interactions of the actors and possible outcomes.

It refers to the same actors as its parent use-case, but because it specifies specific preconditions, the details of the interactions are distinct from other scenarios derived from the same use-case.

An instance of a scenario executed by a benchmark test is referred to as a scenario attempt.

4.2 Benchmark test

A benchmark by definition measures the behavior of a population of users. To accomplish this, the behaviors of individual users must be aggregated into input traffic to the SUT. The input traffic must be realistic, in the sense that a population of users would perform such actions in the real world, and in the sense that statistical variation in user behavior is similar to statistical variation that would occur in the real world.

The top level of this part of the model is a *benchmark test*. A benchmark test defines a *preamble*, the sequence of actions required to initialize a test system and SUT to perform a benchmark; a *traffic set*, a set of scenarios that simulated users perform during the test procedure, together with a relatively frequency with which the scenarios occur during the test procedure; an *arrival distribution*, which describes the arrival rate of occurrences of scenarios from the traffic set; and the *traffic-time profile*, which describes the evolution of the average arrival rate as a function of time over the duration of the test procedure.

During the test procedure, scenarios are selected for execution; the time in between selection and execution is determined by the arrival distribution. The arrival distribution is parameterized by the average arrival rate.

During the test procedure, the scenario arrival rate changes over time according to the traffic-time profile.

The test procedure is carried out as follows:

- The test system first performs the preamble, during which any actions required to initialize it and the SUT are carried out. These actions generally include loading a subscriber base with subscriber data, performing transactions on the subscriber base to randomize the data, and cause the SUT to have "packets in flight" in its internal queues, to make its state approximate the case in which it had been running in a real-world deployment for some extended amount of time.
- The test system sets the initial average arrival rate to the initial value specified by the traffic-time profile. The test system delays for a random interval (calculated by the arrival distribution to achieve the average arrival rate), and then randomly selects a scenario from the traffic set with a probability equal to the scenario percent of system load. This scenario starts to run.
- As time elapses during the test procedure, the profile will change by the SAPS increase amount. When the this value changes, the interarrival time of scenario selection (and hence system load) will change.
- When the entire traffic-time profile has been performed and the total time of the test procedure has elapsed, the test system stops sending scenarios for execution. When the test system completes all currently executing scenarios, the test procedure terminates.

4.2.1 Traffic set

The traffic set is a collection of scenarios which are determined to be likely to co-occur in a real-world scenario. The scenarios need not come from the same use-case. Within a traffic set, each scenario has an associated relative occurrence frequency, interpreted as the probability with which it would occur in the course of the test procedure.

4.2.2 Background load

Background load is a workload presented to the SUT in order to consume its resources. It may consist of a stream of traffic presented to the SUT by an external system apart from the test system; or it may be a workload presented to the processing elements, network, or storage subsystem of the SUT.

The purpose of background traffic is to make possible the measurement of a design objective capacity in an SUT when the capacity of the test system is insufficient to reach the design objective capacity. The use of background load is a pragmatic choice provided for Release 1 of this standard, and remains a study item for future releases.

If a benchmark test is published in which background load is used, then the following requirements apply:

- The hardware used to generate the background load must be fully specified. If the background load is generated by software running directly on the SUT, then the components of the SUT on which the background load is executed must be fully specified.
- The software used to generate the background load must be provided in source form, including make files and any other configuration files required to compile the software.

4.2.3 Arrival distribution

The arrival distribution is a probability model describing the random scenario attempt arrival rate during a benchmark test.

4.2.4 Traffic-time profile

The traffic-time profile is a function describing the average scenario attempt arrival rate as a function of elapsed time during a benchmark test. A traffic-time profile should be chosen in such a manner that, for a given scenario attempt arrival rate, sufficient samples are generated that metrics can be collected with an appropriate confidence bound. For this reason, traffic-time profiles generally follow a "stairstep" pattern, with the width of the stairstep chosen to collect sufficient samples at a constant average scenario arrival rate; the height of the stairstep chosen to provide an appropriate upper and lower bound on metrics such as design objective capacity; and the number of stairsteps chosen to collect data over an appropriate operating range of the benchmark test.

4.2.5 Test parameters

The benchmark test parameters are used to control the behavior of the test script. The data elements required to configure the test system are listed in the table below.

Table 1 is a non-exhaustive list of test parameters defined for the benchmark standard. The list is expected to grow over time, as additional subsystems and system configurations are developed.

Table 1: Test parameters

Parameter	Type	Description
PX_StirTime	float	Amount of time that a system load is presented to a SUT at the beginning of a test after the SUT has been provisioned but before measurements are collected. During this time interval, the database records are said to be "stirred".
PX_SAPSIncreaseAmount	integer	The amount by which the scenario arrival rate is increased.
PX_TotalProvisionedSubscribers	integer	The number of simulated subscribers provisioned in the network.
PX_PercentRegisteredSubscribers	integer	The average percentage of simulated subscribers that are registered simultaneously.
PX_PercentRoamingSubscribers	integer	The average percentage of simulated subscribers that are roaming (i.e. register in a non-local network).
PX_Simulated Subscriber Registration Parameters	integer	Parameters and distributions of the probabilistic model of simulated subscriber operation, as defined in TS 186 008-2 [3].
PX_StepNumber	integer	The number of steps in a benchmark test
PX_StepTime	float	The amount of time that a test is executed at a given system load (a test step) before incrementing the load.
PX_StepTransientTime	float	Interval, from the beginning of a step, during which scenario attempts are not counted.
PX_System Load	integer	Rate at which scenario attempts arrive.

Figure 3: Void

4.3 Benchmark report

A test report is a document, with accompanying data files, that provides a full description of an execution of a benchmark test on a test system. The SUT and test system, as well as their parameters, are described in sufficient detail that an independent test site can replicate the test. The results of the test include data, represented as charts and data sets, depicting the behavior of the SUT over the elapsed time of the test; reports of the relevant metrics that are conventionally used to compare benchmark results of differing SUTs; and a full description of other observations and exceptions noted during the test.

An illustrative example of a benchmark report is presented in annex A. The following clauses describe the elements of a benchmark report.

4.3.1 Disclaimer

This element is narrative text setting forth any certifications of the benchmark test by the testing site, or any disclaimers of certification.

4.3.2 SUT configuration and parameters

This element describes the exact inventory of hardware and software of which the SUT is constructed, a complete description of its configuration and parameter settings as well as characteristics of the interfaces that connect to the test system (like interface bandwidth and latency, interface security characteristics). This description will be different for different SUTs, as the architecture of different SUTs are different from one another, but must be documented in sufficient detail that an independent test site could approach the suppliers of the hardware and software elements and obtain a duplicate system.

4.3.3 Test system configuration and parameters

This element has three sub-elements:

- An exact inventory of hardware and software of which the test system is constructed, and an exact description of its configuration. This description will be different for different test systems, but must be documented in sufficient detail that an independent test site could approach the suppliers of the hardware and software elements and obtain a duplicate system.
- An exact description of the benchmark test run, as documented in clause 4.2. This description must be documented in sufficient detail that an independent test site, using a qualified test system and this description, could duplicate the benchmark result (to within the statistical precision of the measurements) on the same SUT.
- If a background workload was employed in the benchmark test, an exact description of the system used to generate the background load, including a listing of code, is provided.

4.3.4 Report of metrics

The report of metrics contains the following elements:

- Report of Design Objective Capacity (DOC), the primary "comparison" metric reported by this benchmark.
- Reports of additional metrics corresponding to the traffic set, as time series depicted in charts, and with average and standard deviation statistics included.
- Data sets, provided as annexes (possibly as computer-readable files) corresponding to the metrics.

4.3.5 Observations and interpretations

This element has two sub-elements:

- Narrative text, possibly automatically generated by the test system, providing an explanation and guide of how to read the benchmark report.
- Descriptions of any unexpected behavior exhibited by either the test system or SUT, irrespective of whether or not it indicates a deviation from the standard.

4.3.6 Exceptions

This element has two sub-elements:

- A narrative text describing any deviations of the SUT, the test system, or the benchmark test from the standard.
- A description of any errors that occurred during the execution of the benchmark test.

4.3.7 Relationship between system load and number of subscribers

Ideally, the benchmark information model would start with a *user behavioral model* that describes how an individual user in a subscriber population probabilistically generates scenario attempts. The user behavioral model would drive the model of scenario attempt generation that is implemented by the test system. The entire benchmark information model would be anchored in user behavior, and the size of the user population that can be supported by a given SUT could be measured directly.

Currently, there is no agreed model for the number or type of scenarios generated by an individual user. Models used in industry include:

- Assumption that each user of a network makes a certain number of scenario attempts during a busy hour. The number of attempts is fixed by a parameter.

- Assumption that a specific percentage of a user population is active, and that each of the active users makes scenario attempts at a specific rate. The active population percentage, and the scenario attempt rate, is fixed.
- Under the above models, an individual user makes different types of scenarios at different rates.
- Under the above models, an individual user may engage in more than one scenario attempt simultaneously.

Defining a user behavioral model is therefore a subject of further study. This has two consequences for the present release of this multi-part deliverable.

The first consequence concerns the selection of users and scenario attempts. In the ideal case, a user would be selected randomly (according to the user behavioral model), and then the scenario attempt would be selected, probabilistically, but contingent upon the selection of user, again based on the user behavioral model. In the current release of the standard, users are divided into two groups - registered and not registered. For scenarios that assume the user is registered (e.g. successful registration attempt), the user are selected from the not registered group. Correspondingly, for scenarios that assume the user is registered (e.g. successful de-registration attempt) the user is selected from the not registered group. Aside from this state information, the test scenario attempt type and user are selected randomly, but independently from, one another; the information associated with the selected user is used to fill fields of the messages generated by the scenario attempt.

The second consequence concerns the procedure to estimate the size of the user population that can be supported by the SUT. This present document recommends that a test system implementation that wishes to report the size of user population that the SUT can support, should include an additional clause in which the user behavioral model employed is specified.

For example, a benchmark report may define a user behavioral model as:

- The benchmark test assumes that each user served by the SUT makes a single scenario attempt during a busy hour. The number of users supported by the SUT is computed by multiplying the design objective capacity, in scenario attempts per second, by 3 600 (the number of seconds in an hour), and reporting that result as the number of users supported.

NOTE: There is also a relationship between system load and subscriber behavior which may lead to feedback loops which may impact the offered load (e.g. the load could be momentarily increased by scenario re-attempts). Such behavioral model is not considered in the present document.

5 System Under Test (SUT)

Figure 4 depicts the IMS Reference Architecture. The components of the architecture are the primary building blocks, which are either defined by the IMS standard, or defined by external standards and referenced by IMS. The links between the primary building block represent reference points over which the building blocks communicate with each other.

The reference architecture is a logical architecture; no mapping of functional elements to hardware or software component is mandated. And conversely, IMS products as deployed in the real world do not factor neatly into the elements of the reference architecture, which complicates the process of comparing similar products with a benchmark. The problem can be simplified by observing that there are classes of products which have common subsets of reference architecture elements. For classes defined in this manner, common sets of benchmarks can be defined. The classes defined in this manner are called IMS subsystems, or simply subsystems, and are expected to grow over time, as the IMS marketplace becomes richer. TS 186 008-2 [3] of this standard maintains a list and description of subsystems.

In order to proceed from a subsystem description to a benchmark test, a complete description of all aspects of the subsystem relevant to the performance of the benchmark must be present; this is referred to as the system configuration, or SUT configuration. This description starts with an enumeration of the elements of the reference architecture, and an enumeration of all reference points that are external to the subsystem (i.e. reference points between elements within the subsystem are "internal"). However, the configuration requires a specification of the hardware elements (e.g. servers, CPUs, network configuration and bandwidth) and software elements (e.g. operating system, database system), because even though the metrics reported by the benchmark tests are measured with respect to such metrics as total round-trip delay of messages and system capacity, the behavior of IMS systems is still an area of study and a thorough understanding of how CPU and network bandwidth utilization behaves during operation is of interest.

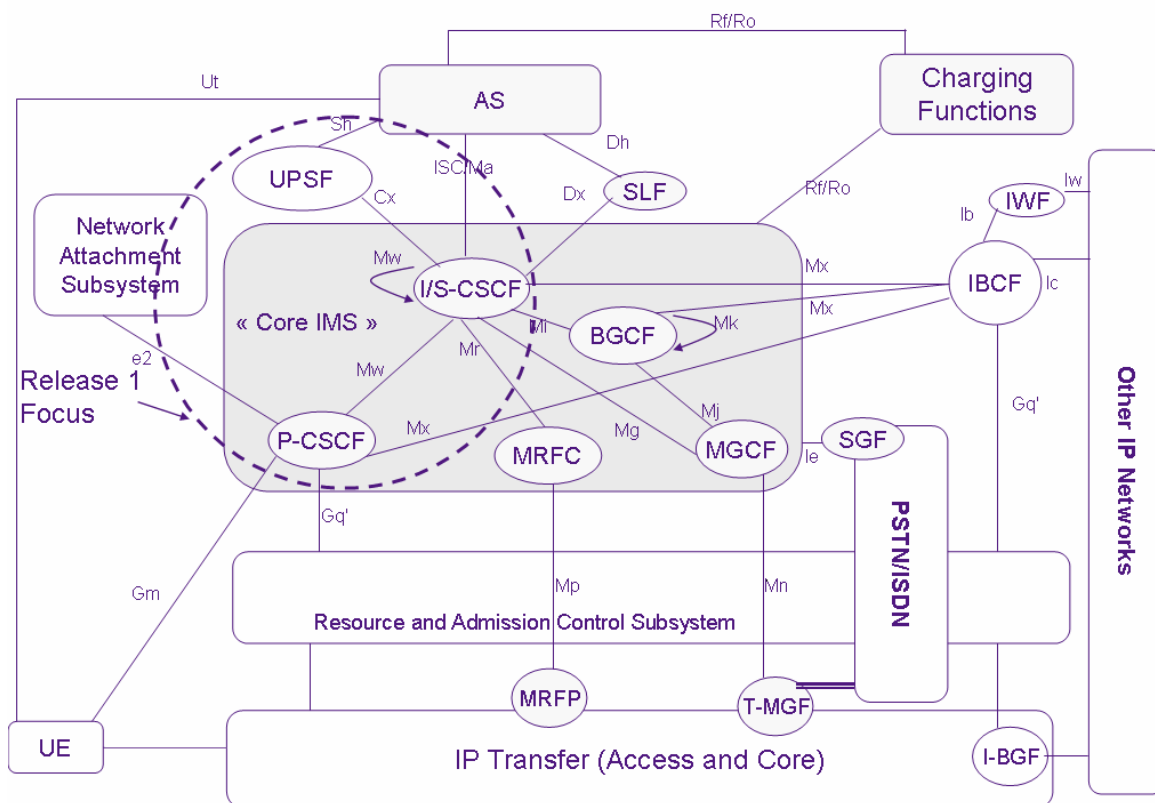


Figure 4: IMS reference architecture (from [1])

5.1 General SUT requirements

An SUT uses the following information elements:

- Subscriber Base: a specification of the subscriber records to be loaded into the HSS/UPSF.
- System Load: a specification of the procedure to be used to generate the system load for each of the SUT interfaces.

5.2 Creation of subscriber base data

The procedure for generating a subscriber base for a benchmark test is presented in TS 186 008-2 [3] of this standard.

The subscriber base is the data-set required to configure the SUT in order to execute a benchmark test. Using the same data, the test system should be able to generate correct traffic.

The present document does not try to specify a complete set, but rather just the subset that will ensure comparable results. All other provisioning information required for correct configuration of a SUT is to be set at the discretion of the SUT provider.

One requirement for a fair benchmark is that the input data is similar for all test-runs. In order for this to happen we have two choices:

- a) Provide data for the subscriber base. However, because we have to ensure scalability for the benchmark, this solution is not feasible.
- b) Provide rules to generate this data and data generators. Algorithms using random generators will be avoided for data that could possibly influence the results.

6 Test system

The test system is used to generate the appropriate load on the SUT. The present document does not mandate any specific test system to be used, although the details of the test system must be reported in the benchmark report.

Figure 5 depicts test system connections and interactions with an SUT. The test system has three main functions:

- **Traffic generation:** the test system must be able to execute use-cases' scenarios following the traffic-time profile. It must also be able to reproduce the appropriate traffic set (a mix of scenarios with a weight for each of them).
- **Network emulation:** optionally, network characteristics on the different interfaces must be emulated by the test system. This includes network bandwidth, latency and error rate. Those characteristics are to be set separately for each direction so that non-symmetric interfaces can be emulated (e.g. up and down bandwidth on a DSL link).
- **Synchronization:** in the case where protocol information elements must be passed between an SUT interface and another and the test system is different for the interfaces, a synchronization mechanism must exist to pass those information elements between the test systems.

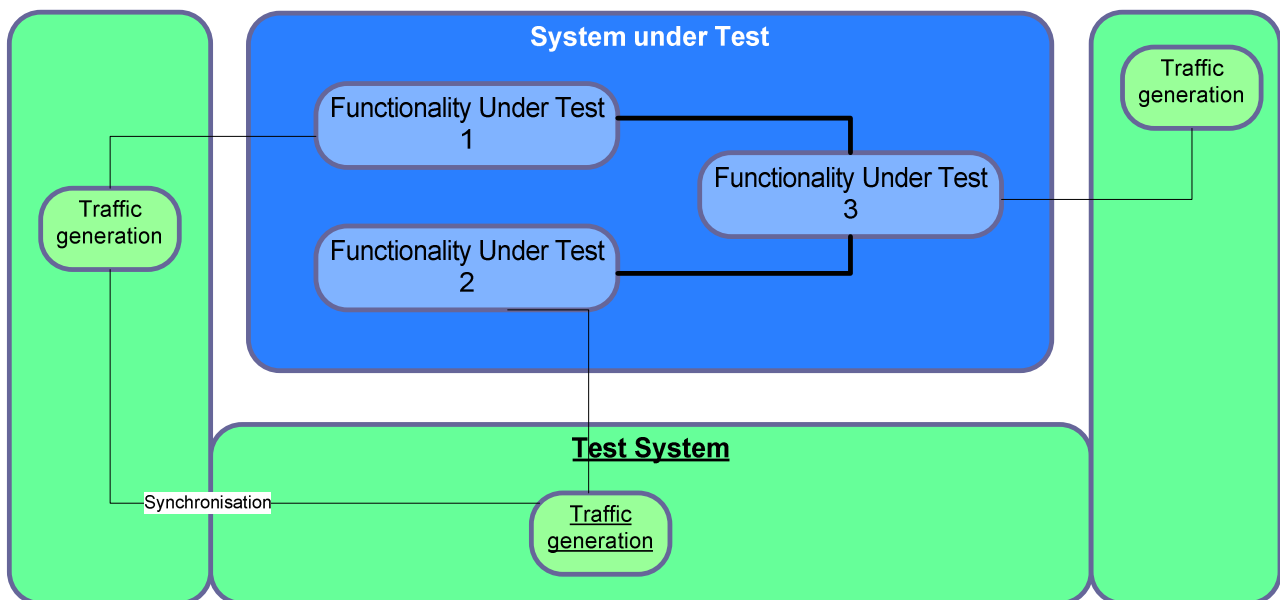


Figure 5: Test System and SUT interactions

7 Benchmark metrics

The metrics reported by a benchmark test may be measured in real time during the execution of the test, or may be computed after completion of the test from event logs collected during the execution.

The metrics are listed in this clause so that they can be selected according to the benchmark to execute. The following information is provided for each metric: short name, full name, description, unit, means to collect the metric, reason to select the metric, acceptance criteria recommendation.

Table 2: Benchmark metrics

Short Name	Full Name	Description	Unit	Means to collect	Purpose	Acceptance criteria
SAPS (see note 1)	Scenario Attempts Per Second	The average rate in a one second period at which scenarios are attempted (not necessary successful)	N/A	On the Test System: as reported by the Test System On the System Under Test: Measured on the reference point with a network analyzer (non intrusive)	Provides a way to define or characterize the load on the SUT	N/A
TRT*	Transaction Response Time*	Defined as the time elapsed from the first message sent to initiate the transaction until the message that ends the transaction is received	ms	Same as SAPS	Measure the processing time of transactions	Scenario dependent - see TS 186 008-2 [3] for details of this family of metrics
CPU	CPU usage ratio	The ratio of used CPU to the total CPU available. This includes all CPUs available for the processing. This ratio is typically expressed as a percentage	%	Measured by a monitoring application like "top", or collected by SNMP MIB queries	Verify that CPU is not overloaded, or extrapolate benchmark results to different CPU	
MEM	Memory usage ration	The ratio of used memory to the total memory available	%	Measured by a monitoring application like "top", or collected by SNMP MIB queries	Verify that memory usage is stable over time	
RETR	Retransmission Rate	Applies to UDP transport. The retransmission rate is the ratio between the number of retransmission (a message can be retransmitted several times) and the number of messages sent	%	Same as SAPS	Unveil overload of the system under test	
SIMS	Simultaneous Scenarios	Number of scenarios that are active at the same time		Same as SAPS	Evaluate the impact parallel scenarios	
%IHS	Percent Inadequately Handled Scenarios	The ratio of inadequately handled scenarios to the total number of attempted scenarios (see below)	%		Provide the Design Objective Capacity	- Under nominal load: $\leq 0,1$ % - Under stress condition (overload): ≤ 10 % (see note 2)
NOTE 1: The "SAPS" metric is related to load, rather than to performance.						
NOTE 2: Those values are network dependent. A particular network might require much tighter constraints, like $< 0,001$ % under nominal load.						

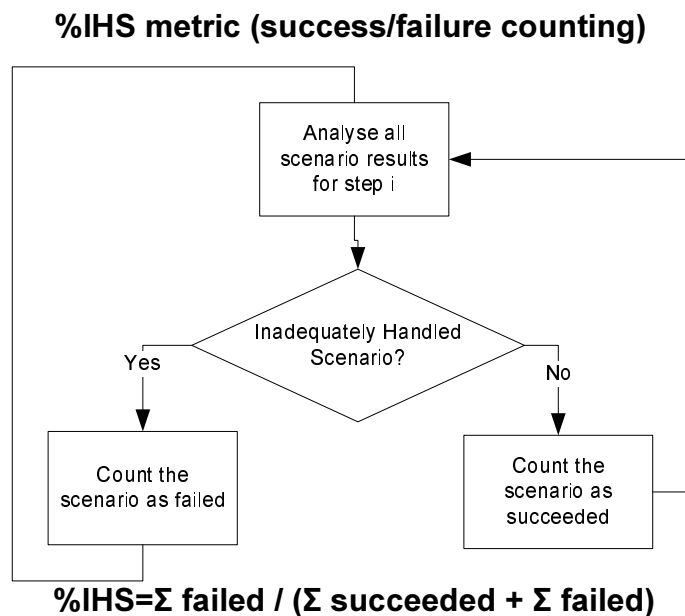


Figure 6: %IHS metric (success/failure counting)

The %IHS metric, which is ultimately used to find the Design Objective Capacity, is computed by dividing the number scenarios that failed by the number of scenarios that succeeded, during a step "i".

A scenario is counted as failed when either:

- The message flow was not respected.
- The message flow was stuck (timeout).
- One of the acceptance criteria for the scenario is not met.

If one of those three conditions is true, then the scenario is considered to be an "Inadequately Handled Scenario".

Annex A (informative): Illustrative example of benchmark report

This example focuses on the Registration scenario and measures SAPS, or Scenario Attempts Per Second, with 100 % of the scenarios being registration attempts. Load is started at 40 SAPS, and increases (steps) 4 times at 10 SAPS per step to the final value of 80 SAPS.

The report was generated by a prototype tool that analyzes the log file generated by a TTCN-3 test system. In its current release, the terms used in the report are not completely aligned with the Standard; it is provided in order to help to create an intuitive impression of the information provided by the benchmark.

A.1 Benchmark

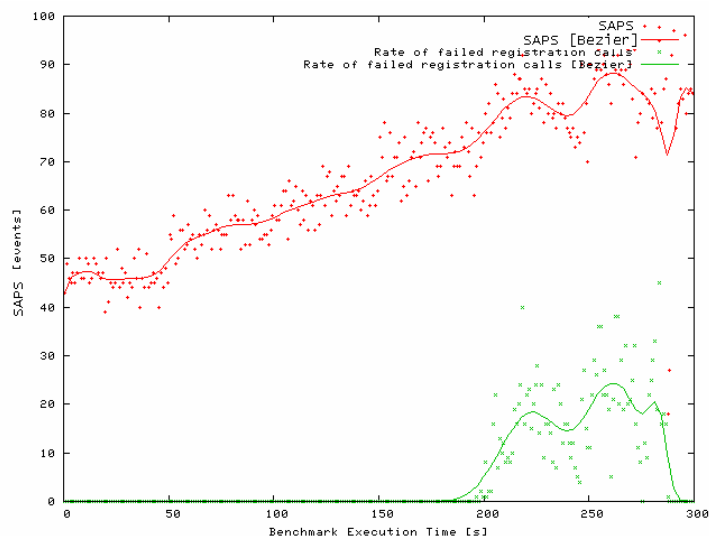
PX_StirTime	60 s
PX_SAPSIncreaseAmount	10
PX_TotalProvisionedSubscribers	3 800
PX_PercentRegisteredSubscribers	100 %
PX_PercentRoamingSubscribers	0 %
PX_StepNumber	4
PX_StepTime	60 s
PX_StepTransientTime	0
PX_SystemLoad	40

A.2 Benchmark results

- 1) Overall Traffic (Time Series) (overall).
- 2) Registration Traffic (Time Series) (overall).
- 3) Average Duration of Completed Registrations (Time Series) (overall).
- 4) Histogram of Delays Between Consecutive REGISTER Requests (Histogram) (overall).

A.2.1 Registration traffic (time series) (overall)

This graph shows the registration traffic of the test system.



Rate of initiated registrations

This line represents the call rate for the registrations sent during the benchmark time.

	Maximum	Minimum	Average	Variance	Standard Deviation	Median	Mode
Initiated Registrations Per Second [events]	98,00	18,00	66,68	225,80	15,03	66,00	63,00

The maximum of 98,00 events occurred first at 276,50 s.

The minimum of 18,00 events occurred first at 287,50 s.

The first point on the graph is at 0,50 s. The last point is at 299,50 s.

The average value is 66,68 events.

The degree of homogeneity is shown by the variance (225,80) and/or standard deviation (15,03 events); small values indicate that data is clustered near the average value.

The median is 66,00 events; half of the points are above and half below this value.

The most frequent value encountered is 63,00 events.

Rate of failed registrations

This line represents the fail rate for the registrations during the benchmark time.

	Maximum	Minimum	Average	Variance	Standard Deviation	Median	Mode
Initiated Registrations Per Second [events]	45,00	0,00	5,24	92,94	9,64	0,00	0,00

The maximum of 45,00 events occurred first at 283,50 s.

The minimum of 0,00 events occurred first at 0,50 s.

The first point on the graph is at 0,50 s. The last point is at 299,50 s.

The average value is 5,24 events.

The degree of homogeneity is shown by the variance (92,94) and/or standard deviation (9,64 events); small values indicate that data is clustered near the average value.

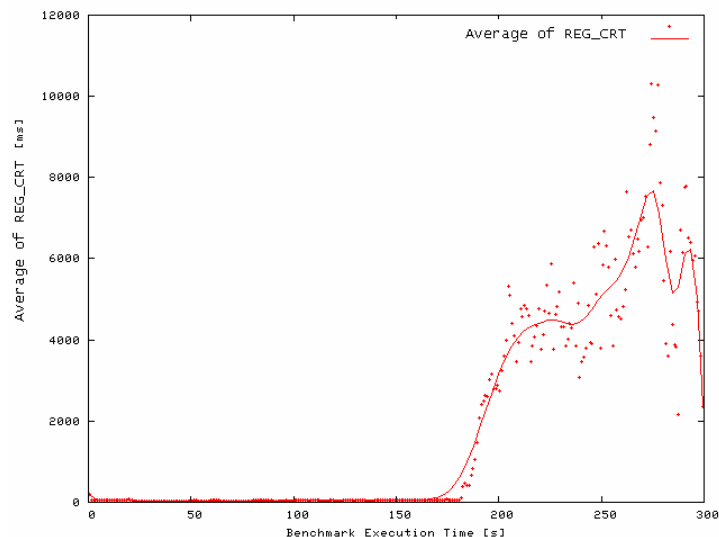
The median is 0,00 events; half of the points are above and half below this value.

The most frequent value encountered is 0,00 events.

Error above the 0,03 % of the step rate first occurs at 198,50 s; the value is 2,0 events.

A.2.2 Average duration of completed registrations (time series) (overall)

This graph presents the average duration of the successful registrations during the benchmark time. It is useful for emphasizing the evolution of the SUT behavior and its response times during the entire benchmark.



Average duration of registrations

Average duration of all the registrations initiated during each sub-interval.

	Maximum	Minimum	Average	Variance	Standard Deviation	Median	Mode
Average Delay of Registrations [ms]	10 323,22	24,07	1 870,07	6 674 071,23	2 583,42	57,60	29,00

The maximum of 10 323,22 ms occurred first at 274,50 s.

The minimum of 24,07 ms occurred first at 22,50 s.

The first point on the graph is at 0,50 s. The last point is at 299,50 s.

The average value is 1 870,07 ms.

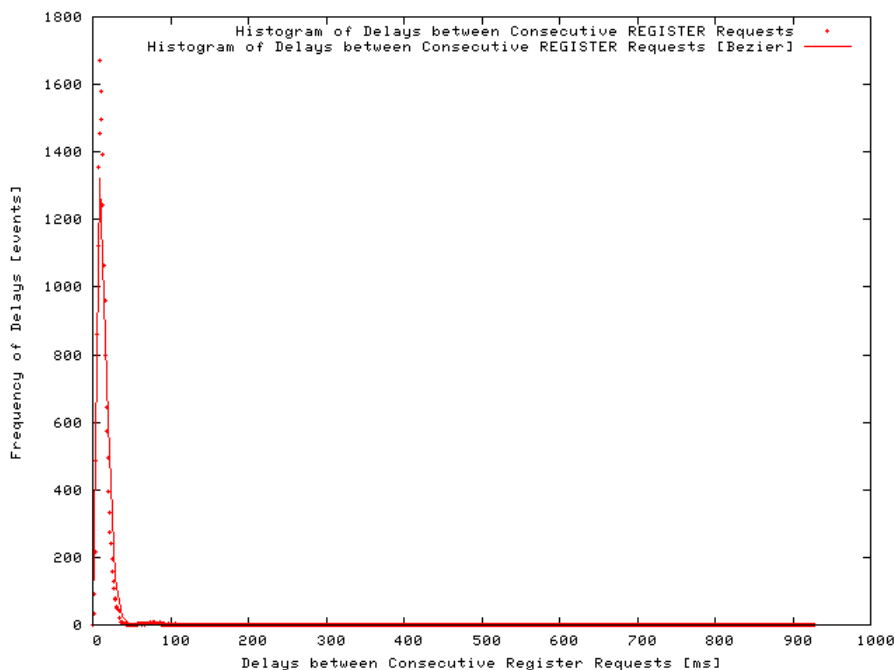
The degree of homogeneity is shown by the variance (6 674 071,23) and/or standard deviation (2 583,42 ms); small values indicate that data is clustered near the average value.

The median is 57,60 ms; half of the points are above and half below this value.

The most frequent value encountered is 29,00 ms.

A.2.3 Histogram of delays between consecutive REGISTER requests (Histogram) (overall)

This graph shows the histogram of the delays between sending successive REGISTER requests. This can be helpful in checking the pattern by which the requests are transmitted (e.g. Poisson).



Histogram of Delays between Consecutive REGISTER Requests

	Maximum	Minimum	Average	Variance	Standard Deviation	Median	Mode
Frequency of Delays [events]	1 673,00	0,00	21,56	22 678,28	150,59	0,00	0,00

The maximum of 1 673,00 events occurred first at 9,50 ms.

The minimum of 0,00 events occurred first at 45,50 ms.

The first point on the graph is at 0,50 ms. The last point is at 927,50 ms.

The average value is 21,56 events.

The degree of homogeneity is shown by the variance (22 678,28) and/or standard deviation (150,59 events); small values indicate that data is clustered near the average value.

The median is 0,00 events; half of the points are above and half below this value.

The most frequent value encountered is 0,00 events.

Annex B (informative): TTCN-3 based test system sample

A TTCN-3 test system is contained in archive ts_18600801v010101p0.zip which accompanies the present document, as an example and for illustration purposes. This implementation has been used to validate the present benchmark specification. It implements registration, session setup and tear down and messaging use cases.

Annex C (informative): Bibliography

IETF RFC 3310 (September 2002): "Hypertext Transfer Protocol (HTTP) Digest Authentication Using Authentication and Key Agreement (AKA)".

IETF RFC 3840 (August 2004): "Indicating User Agent Capabilities in the Session Initialization Protocol (SIP)".

ETSI TR 121 905 (V7.0.0): "Digital cellular communications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Vocabulary for 3GPP Specifications (3GPP TR 21.905 version 7.0.0 Release 7)".

ETSI TS 183 041 (V1.1.1): "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Messaging service using the IP Multimedia (IM) Core Network (CN) subsystem; Stage 3: Protocol specifications [Endorsement of 3GPP TS 24.247 Release 6]".

ETSI TS 123 228: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); IP Multimedia Subsystem (IMS); Stage 2 (3GPP TS 23.228 Release 6)".

ETSI TS 124 247: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Messaging service using the IP Multimedia (IM) Core Network (CN) subsystem; Stage 3 (3GPP TS 24.247)".

History

Document history		
V1.1.1	October 2007	Publication