ETSI GS NFV-INF 010 V1.1.1 (2014-12)

Network Functions Virtualisation (NFV); Service Quality Metrics

Disclaimer

This document has been produced and approved by the Network Functions Virtualisation (NFV) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGS/NFV-INF010

Keywords
measurement, NFV, quality

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

Contents

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://ipr.etsi.org).

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 Group Specification (GS) has been produced by ETSI Industry Specification Group (ISG) Network Functions Virtualisation (NFV).

The present document deals with specific aspects of Service Quality Metrics in the context of Network Function Virtualisation.

Infrastructure Architecture Document		**Document #**
Overview		GS NFV INF 001
Illustrative Use Cases for the NFV Infrastructure		GS NFV INF 002
Architecture of the Infrastructure Domains	Compute Domain	GS NFV INF 003
	Hypervisor Domain	GS NFV INF 004
	Infrastructure Network Domain	GS NFV INF 005
Architectural Methodology	Interfaces and Abstraction	GS NFV INF 007
Service Quality Metrics		GS NFV INF 010

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**may not**", "**need**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

As shown in figure 1, the service quality delivered by a VNF instance to end users is dependent on the service quality of the compute, network and other resources delivered by NFV infrastructure, VIM, VNFM and NFVO to the VNF instance. Objective and quantitative metrics for the service delivered by NFV infrastructure and orchestration to NFV consumers enables:

- Better engineering of VNF user service quality.
- Efficient fault localization and mitigation.
- Faster identification of true root cause of service impairment so proper corrective actions can be taken promptly.

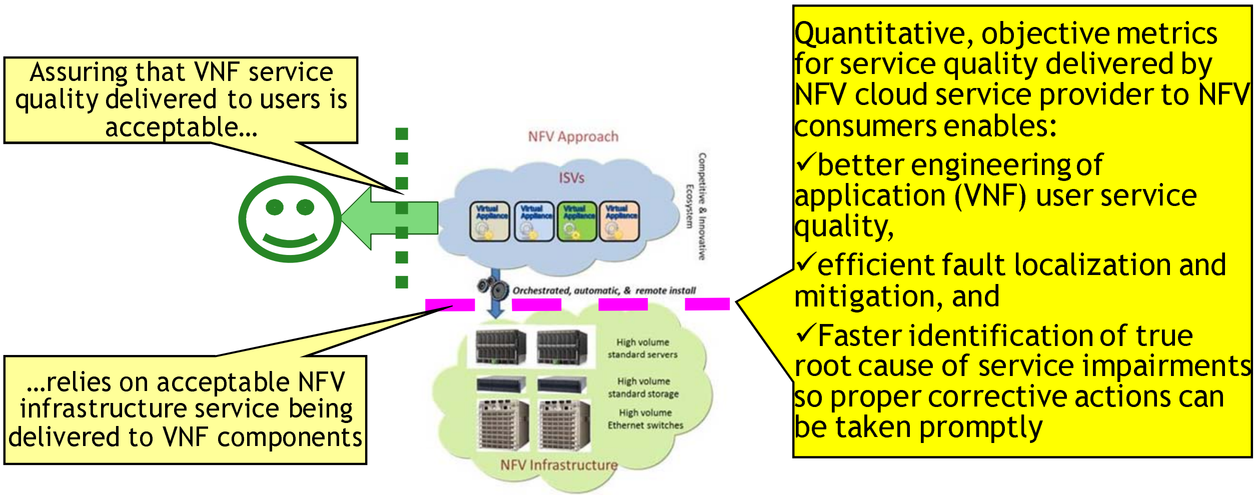

Figure 1: Purpose of NFV Service Quality Metrics

Figure 2 illustrates the service boundary characterized by these metrics between canonical NFV consumers who operate VNFs and NFVI service providers who operate NFV infrastructure and supporting systems. Objective metrics of the service quality delivered by the NFV infrastructure service provider to the NFV consumer's VNFs enable quantitative discussions and agreements for the objectives of NFV service quality to assure that the NFV consumer's VNF instances deliver acceptable service quality to end users.

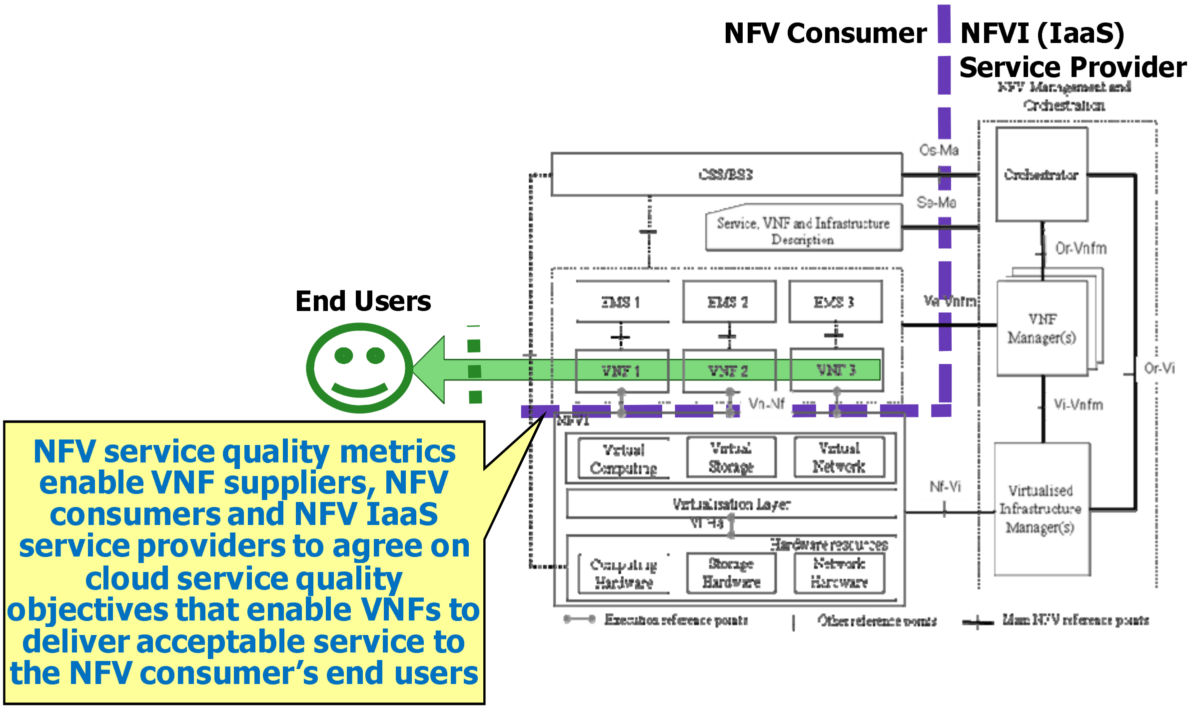

Figure 2: Service Quality Metrics in the Context of NFV Reference Architecture

Both new and familiar metrics will be needed to complete the quantification of service quality. The present document examines the properties of the virtual infrastructure and describes metrics relevant and useful to virtualised functions.

1 Scope

The present document enumerates metrics for NFV infrastructure, management and orchestration service qualities that can impact the end user service qualities delivered by VNF instances hosted on NFV infrastructure. These service quality metrics cover both direct service impairments, such as IP packets lost by NFV virtual networking which impacts end user service latency or quality of experience, and indirect service quality risks, such as NFV management and orchestration failing to continuously and rigorously enforce all anti-affinity rules which increases the risk of an infrastructure failure causing unacceptable VNF user service impact. Performance relationships exist between the metrics described in this document and in other specifications such as [i.5].

The present document does ***not*** consider:

1) Units of measurement for reporting, such as whether VM premature release rates should be expressed as hourly rate (e.g. 0,0001 premature VM release events per hour), annualized rate (e.g. 0,88 premature VM release events per year), hours between events (e.g. 10 000 hour mean time between premature release events), or events per other unit of time (e.g. 100 000 FITs, meaning 100 000 premature release events in one billion hours of operation).

2) Methods of Measurement which stipulate exactly how metrics will be measured.

3) Rigorous counting and exclusion rules, like the precise details given in the TL 9000 Measurements Handbook [i.13].

4) Metrics that do not directly or indirectly impact VNF user service quality, like power efficiency.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

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

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

The following referenced documents are necessary for the application of the present document.

[1] ETSI GS NFV-INF 001: "Network Functions Virtualisation (NFV); Infrastructure Overview".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

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

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] NIST Special Publication 800-145 (September 2011): "The NIST Definitions of Cloud Computing", Peter Mell and Timothy Grance, US National Institute of Standards and Technology.

[i.2] IETF RFC 2330: "Framework for IP Performance Metrics".

[i.3] Wiley-IEEE Press, 2013: "Service Quality of Cloud-Based Applications," Eric Bauer and Randee Adams.

[i.4] ETSI GS NFV-MAN 01: "Network Functions Virtualisation (NFV); Management and Orchestration".

[i.5] draft-ietf-ippm-model-based-metrics-02 (work in progress) (February 2014): "Model Based Bulk Performance Metrics", M. Mathis and A. Morton.

[i.6] ETSI GS NFV-PER 001: "Network Functions Virtualisation (NFV); NFV Performance & Portability Best Practises".

[i.7] IETF RFC 6390: "Guidelines for Considering New Performance Metric Development".

[i.8] ISO/IEC 15939:2007: "Systems and software engine-ing -- Measurement process".

[i.9] NIST draft Cloud Service Metric Description v2.

[i.10] Recommendation ITU-T M.3341: "Requirements for QoS/SLA management over the TMN X-interface for IP-based services".

[i.11] Recommendation ITU-T Y.1543: "Measurements in IP networks for inter-domain performance assessment".

[i.12] Recommendation ITU-T I.356: "B-ISDN ATM layer cell transfer performance".

[i.13] TL 9000 Measurements Handbook, release 5.0, July 2012, QuestForum.

NOTE: Available at http://www.tl9000.org/handbooks/measurements_handbook.html.

3 Definitions and abbreviations

3.1 Definitions

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

derived metric: metric defined on the basis of the values produced by other Metrics

NOTE: An example from packet transfer performance is delay variation, which is based on multiple values produced from measurement of a packet delay metric. (This definition is consistent with IETF RFC 2330 [i.2] and the derived performance parameter in Recommendation ITU-T I.356 [i.12]).

measurement: set of operations having the object of determining a Measured Value or Measurement Result

NOTE: The actual instance or execution of operations leading to a Measured Value. (Based on the definition of Measurement in IETF RFC 6390 [i.7], as cited in ISO/IEC 15939 [i.8]).

measurement point: physical or logical point at which observations are made and to which the measure obtained is related, e.g. a boundary or point of demarcation between domains or functional entities (derived from NIST draft Cloud Service Metric Description v2 [i.9])

NOTE 1: When one or more measurement points are specified along with a performance metric, they define the scope of measurement and should be included with the measurement results for accurate interpretation.

NOTE 2: A point in a system possessing sufficient functionality to provide observations of reference events. (derived from Recommendation ITU-T M.3341 [i.10]) The interface between two communicating entities is often designated as a Measurement Point.

metric: standard definition of a quantity, produced in an assessment of performance and/or reliability of the network, which has an intended utility and is carefully specified to convey the exact meaning of a measured value

NOTE: This definition is consistent with that of Performance Metric in IETF RFC 2330 [i.2] and ETSI GS NFV-PER 001 [i.6].

EXAMPLE: Packet transfer performance or reliability of a network.

parameter: input factor defined as a variable in the definition of a Metric

NOTE: A numerical or other specified factor forming one of a set that fully-defines a Metric or sets the conditions of its operation. Most Parameters do not change the fundamental nature of the metric's definition, but others have substantial influence. All Parameters should be known in order to conduct measurements conforming to a Metric and interpret the results. An example Parameter includes the Measurement Point(s). (derived from IETF work in progress to design a Registry for Performance Metrics, consistent with IETF IPPM Literature)

reference event: transfer of a discrete unit of control or user information encoded in accordance with a specific protocol across a Measurement Point

NOTE: Complementary classes of reference events can sometimes be distinguished: exit and entry events, or start and stop events (derived from Recommendation ITU-T Y.1543 [i.11]).

3.2 Abbreviations

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

CCDF Complementary Cumulative Distribution Function
CPU Central Processing Unit
DOA Dead on Arrival

NOTE: Also referred to as "Out-of-Box" (OOB) failures.

HW&SW Hardware and Software
IETF Internet Engineering Task Force
IP Internet Protocol
MIB Management Information Base
NFV Network Function Virtualisation
NFVI NEFV Infrastructure
NFVO NFV Orchestrator
NIC Network Interface Card
OS Operating System
SLA Service Level Agreement
SLO Service Level Objective
SQM Service Quality Metrics
TcaaS Technology Component offered as-a-Service

NOTE: Like Database-as-a-Service.

UTC Universal Coordinated Time
VIM Virtual Infrastrucutre Management
VM Virtual Machine
VN Virtual Network
VNF Virtualised Network Function
VNFC Virtualised Network Function Component
VNFM VNF Manager

4 NFV Service Quality Metrics Taxonomy

End users experience services delivered by VNF instances, which are implemented via suites of VNFCs working together. The services delivered to end users by individual VNFC instances is dependent on the service quality of the virtual machine instance that hosts the component and the virtual network service that delivers connectivity to other VNFCs. End user service quality of VNFs which use functional blocks or technology components that are offered 'as-a-Service' like Database-as-a-Service or Load-Balancing-as-a-Service are also vulnerable to service quality impairments of those technology components. NFV management and orchestration can also presents subtle risks to VNF service quality if elastic resource growth or repair is slow or faulty, or if the VNF's anti-affinity rules are not strictly and continuously enforced. Figure 3 visualizes the four suites of NFV service quality metrics:

1) Virtual machine service quality metrics.

2) Virtual network service quality metrics.

3) Technology components offered 'as-a-Service' (e.g. Database-as-a-Service) quality metrics.

4) Orchestration service quality metrics.

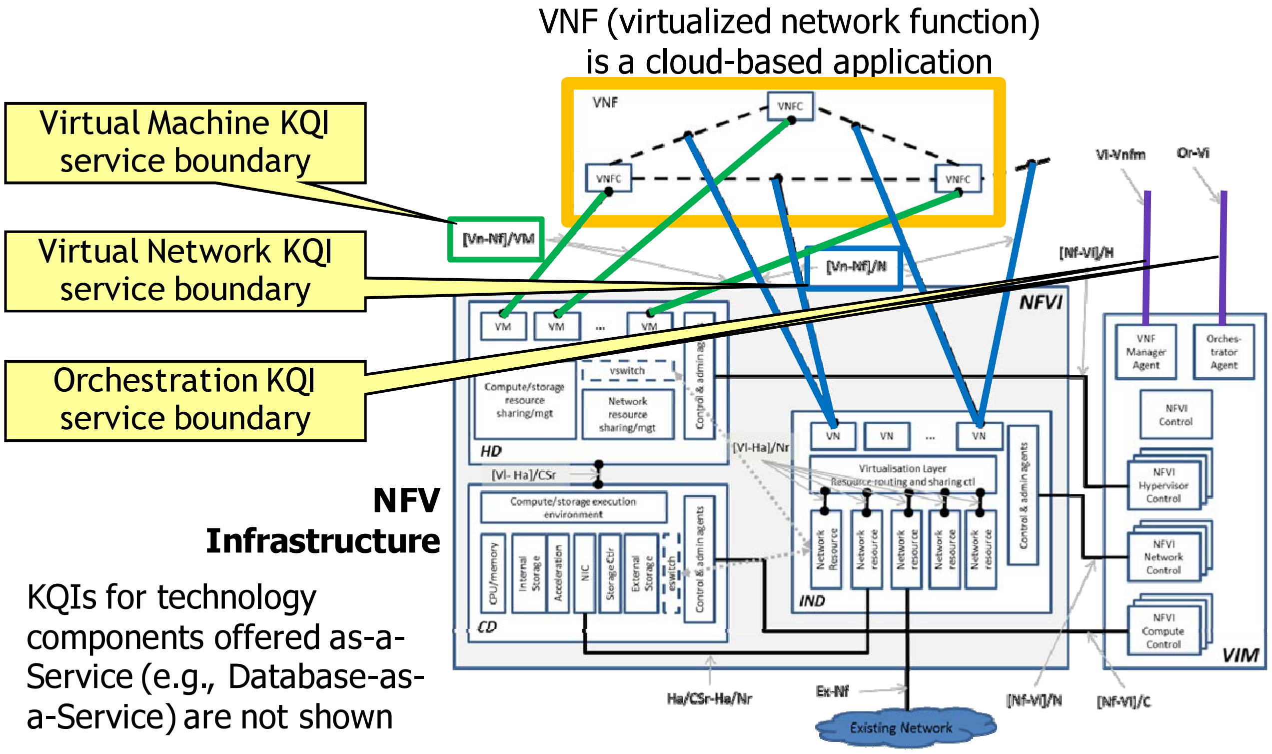

Figure 3: NFV Service Quality Metrics on High Level Overview of NFVI

When assessing performance of the NFV Infrastructure, measurement points at the [Nf-Vi]/* reference points will be critical. Measurement points at other locations in the NFV architecture will necessarily include the performance contribution of management components, such as the Vi-Vnfm and Or-Vi reference points, and these may also prove to be valuable aids in performance management.

The NFV service quality metrics are summarized in table 1. Each cell of the matrix is associated with a Service/life-cycle category (e.g. Orchestration, Operation) and a quality criterion (Speed, Accuracy, Reliability). Some intersections are critical for certain functions or resources while others may be inapplicable or contain secondary metrics. By listing each Quality Metric in its corresponding cell, it is possible to identify overlaps and gaps. The matrix may also facilitate the process to determine which Quality Metrics are the key ones and should be measured, collected, and reported. The Service Quality Metrics are applicable when the entity measured is deemed to be in the Available state (through continuous evaluation of one or more metrics). See [i.3] for consistent definitions of Availability and Reliability.

Table 1: Summary of NFV Service Quality Metrics

Service Metric Category	Speed	Accuracy	Reliability
Orchestration Step 1 (e.g. Resource Allocation, Configuration and Setup)	VM Provisioning Latency	VM Placement Policy Compliance	VM Provisioning Reliability VM Dead-on-Arrival (DOA) Ratio
VirtualMachine operation	VM Stall (event duration and frequency) VM Scheduling Latency	VM Clock Error	VM Premature Release Ratio
Virtual Network Establishment	VN Provisioning Latency	VN Diversity Compliance	VN Provisioning Reliability
Virtual Network operation	Packet Delay Packet Delay Variation (Jitter) Delivered Throughput	Packet Loss Ratio	Network Outage
Orchestration Step 2 (e.g. Resource Release)			Failed VM Release Ratio
Technology Component as-a-Service	TcaaS Service Latency	-	TcaaS Reliability (e.g.defective transaction ratio) TcaaS Outage

The impact of each NFV service quality metric of table 1 can directly or indirectly impair the VNF user service quality as follows:

- **VM provisioning latency** and **VM provisioning reliability** directly impact the time it takes to elastically grow online VNF service capacity or to restore full VNF redundancy (i.e. eliminate simplex exposure) following a failure event.
- **VM Dead on Arrival** (DOA) indirectly impacts the time to elastically grow or repair VNF capacity because a latent VM fault shall be detected and somehow mitigated before needed VNF capacity can enter user service.
- **VM premature release ratio** directly impacts the frequency that VNF service recovery actions (e.g. high availability failovers) shall be taken.
- **VM stall** characterizes disruptions in prompt and continuous execution of VNFC software which impacts the service latency and quality of service enjoyed by end users.
- **VM scheduling latency** characterizes how promptly VNFC software is executed, such as when processing isochronous bearer plane traffic, which impacts the service latency and quality of service enjoyed by end users.
- **VM clock error** characterizes the realtime inaccuracy presented to VNFC software for billing, fault and other records that rely on accurate timestamping.
- **VM placement policy compliance** characterizes how reliably the NFV infrastructure provider continuously and correctly enforces the VNF's anti-affinity rules, thereby minimizing the risk that a single infrastructure failure event will produce an unacceptable VNF user service quality impact.
- **Packet delay** characterizes the incremental user service latency introduced by communications between a VNF's VNFCs, which impacts the service latency and quality of service enjoyed by end users. A key input parameter for any packet transfer metric is the offered load during the measurement.
- **Packet delay variation (jitter)** is a derived metric that characterizes the incremental user service delay variation introduced by instability in communications latency between VNFCs within a VNF, which impacts the service latency and quality of service enjoyed by end users.
- **Delivered Throughput** is a derived metric from the offered load input parameter and other packet transfer performance metrics (loss, delay) measured at that load to characterize the actual capacity of communications between a VNF's VNFCs, and which impacts the quality of service enjoyed by end users (ETSI GS NFV-MAN 001 [i.4]).

- **Packet loss ratio** impacts end user service latency, reliability and quality because lost packets shall be detected, and mitigated via retry, retransmission or concealment, which impacts the service latency and quality of service enjoyed by end users. Lost packets should be assessed at measurement points near the ingress and egress of the network being measured. Again, a key input parameter for any packet transfer metric is the offered load during the measurement.

- **Network outage** loss of virtual network connectivity directly impacts the service latency, quality and availability experienced by end users. Network impairment episodes that persist longer than the VNF's Maximum Acceptable Network Transient Time parameter will prompt highly available VNFs to automatically initiate service recovery actions, up to and including VNF disaster recovery actions.

- **Failed VM Release Ratio** characterizes how reliably the usage records for VM release are recorded.

- **VN provisioning latency** and **VN provisioning reliability** directly contribute to the time needed to create or add VNF service capacity, or to restore full VNF redundancy. These metrics track the time to successfully establish Infrastructure Network connectivity when requested, and the establishment attempts that fail, respectively.

- **VN diversity compliance** characterizes how accurately the NFV Infrastructure Network achieves the requested physical diversity, thereby minimizing the risk that a single infrastructure failure event will produce an unacceptable VNF user service quality impact. Connectivity requests involving VN diversity could have a parameter giving the minimum number of physically diverse paths (2, 3, etc.).

The end user service quality of VNFs that rely on technology components offered as-a-service (e.g. Load-Balancing-as-a-Service, Database-as-a-Service) is directly impacted by:

- **TcaaS service latency** which adds directly to VNF service latency, which impacts the service latency and quality of service enjoyed by end users.

- **TcaaS service reliability** because transaction failures either flow directly back to the end user as errors, or indirectly impact service latency as mitigation actions are executed, which impacts quality of service enjoyed by end users.

- **TcaaS service outage** events directly impact VNF service.

Note that the end user service of different types of VNFs will be sensitive to different types of NFV impairments, and thus different VNFs will be most concerned with different NFV service quality metrics. For example, the service quality experienced by an end user of a bearer plane element like a session border controller is highly influenced by VM stalls, VM scheduling latency and packet loss, while the service quality of an operations support system that relies on Database-as-a-Service is highly influenced by the service reliability and service availability of that Database-as-a-Service offering.

5 Virtual Machine Service Quality Metrics

Figure 4 visualizes the VM lifecycle as a timeline of a single VM instance. The lifecycle begins with a request to the NFV management and orchestration domain to allocate a new VM instance to startup, grow or repair VNF capacity. The response to that request is either a newly allocated VM instance or a failure indication (e.g. resources unavailable). The newly allocated VM instance is then integrated with the VNF. Integration of the newly allocated VM with the VNF instance can either be successful in which case the new VM instance begins serving VNF users at time "***T=0***" or the integration fails, such as because the allocated VM instance was inoperable or "dead-on-arrival". The VM instance's useful life begins at ***T=0*** when it is available to serve VNF users. The useful life ends either with an orderly release when the application domain gracefully releases the VM instance via the management and orchestration domain (e.g. to elastically shrink VNF capacity); alternately, the VM can be prematurely released due to failure or for other reasons (e.g. executing a lawful takedown order or non-payment of bill).

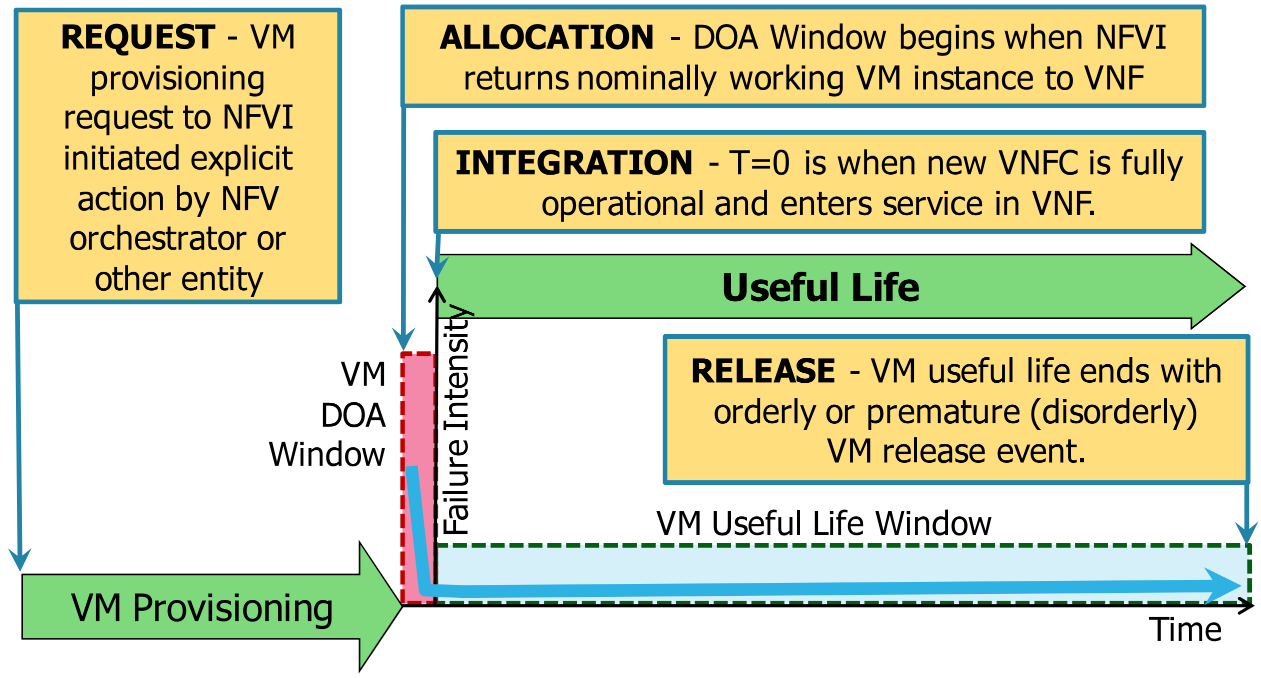

Figure 4: Virtual Machine Lifecycle

VM-related NFV service quality metrics before "T=0" when the VM enters service (i.e. VM Provisioning Latency, VM Provisioning Reliability, VM DOA Ratio) are covered as orchestration service quality metrics. Figure 5 illustrates that the virtual machine service quality metrics (VM stall, VM premature release ratio, VM scheduling latency and VM clock error) apply from "T=0" when the target VM becomes available to serve VNF users until the VM is released in either an orderly or disorderly way.

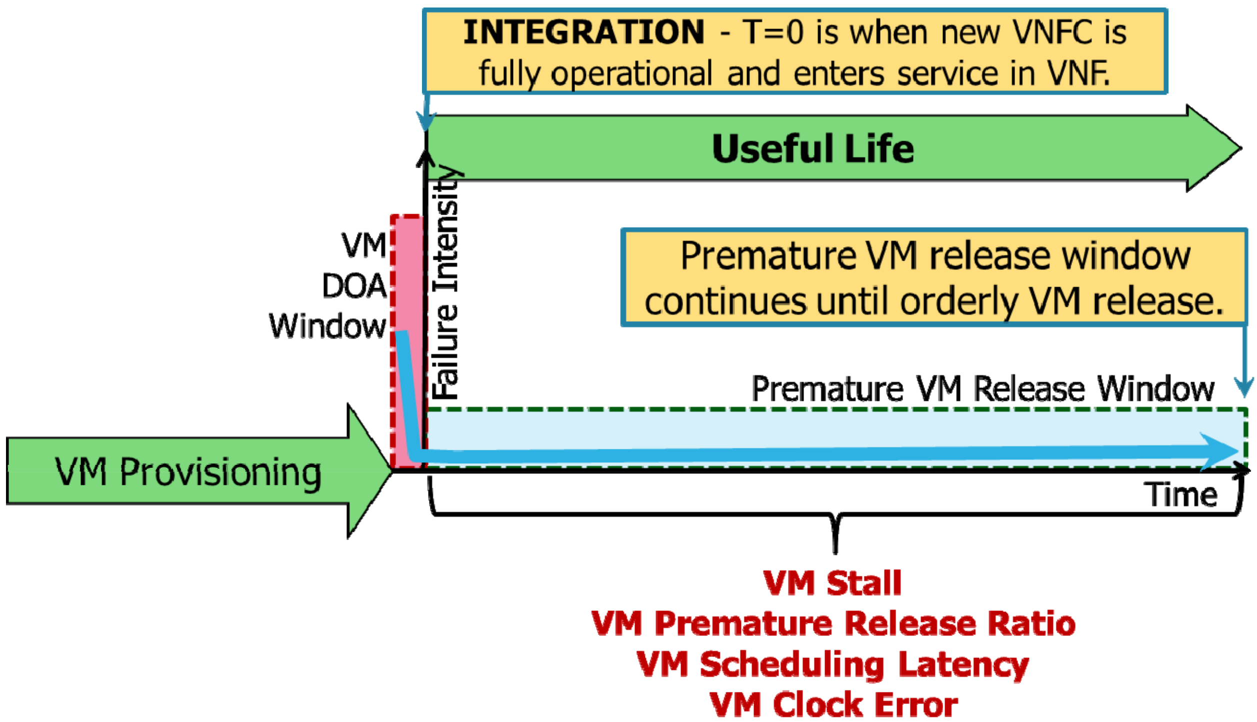

Figure 5: Virtual Machine Service Quality Metrics

5.1 VM Stall

VM instances can briefly cease execution, stall or 'hiccup', for reasons not explicitly requested by the application or cloud consumer, such as "live" VM migration. During the stall event, the application component hosted by the impacted VM instance does not execute, so both new requests and queued work fails to be served for the duration of the stall event, thereby potentially impacting end user service. Individual VM stall events are measured as the elapsed time between when the VM instance ceases to be executed (e.g. when a VM is paused at the start of a 'live' VM migration) to the instance when the VM instance resumes execution (e.g. when VM instance is resumed at the conclusion of a 'live' VM migration event). The primary metrics of VM stall impairments are (1) duration of VM stall event, and (2) frequency of VM stall events. These metrics can be used for, e.g. root cause study as explained below in clause 9.2.

If execution of a VM instance is suspended for longer than the application's maximum acceptable VM stall time, then the VNF's high availability mechanism will activate to failover user service to another VM instance. As shown in figure 6, stall events that exceed the VNF's maximum acceptable stall time are counted as premature VM release events rather than VM stall events.

5.2 Premature VM Release Ratio

Premature VM release ratio measures the intensity of disorderly VM release events. Premature VM release events explicitly include the following disorderly release event causes:

1) NFV infrastructure failure event, like host hardware or hypervisor failure, or power disruption to that host.

2) VM stall events whose duration exceeds the VNF's maximum acceptable VM stall time where the maximum acceptable stall time is an input parameter of the metric.

3) Unavailability or unacceptably poor performance for at least the VNF's maximum acceptable VM stall time for the virtual CPU, virtual NIC or virtual disk resources allocated to the target VM instance.

Premature VM release measurements exclude the following release event causes:

1) VM release requested by VNF or NFV consumer.

2) VM release triggered by execution of lawful takedown order.

3) VM release triggered by non-payment of bill by NFVI consumer, if applicable.

4) VM release triggered by failure to comply with the NFV infrastructure service provider's acceptable use provisions.

Figure 6 juxtaposes VM stall and VM premature release events. An impairment that renders a VM inoperable for a period of no more than the maximum acceptable VM stall time is charged as a VM stall event; if the inoperable period exceeds the maximum acceptable VM stall time then the event is deemed a premature VM release.

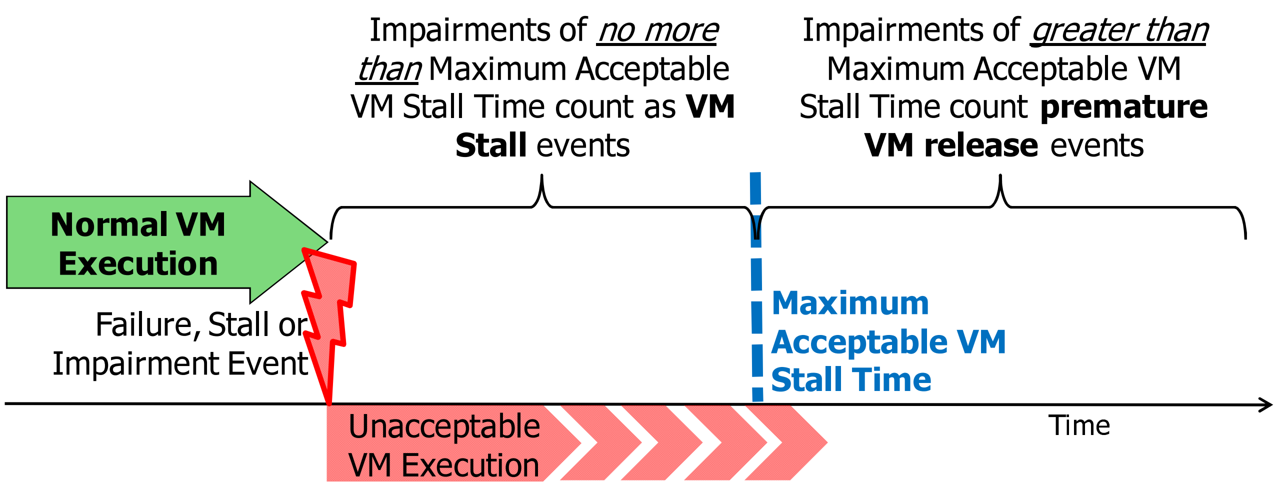

Figure 6: VM Stall and Premature VM Release Impairments

5.3 VM Scheduling Latency

By inserting a hypervisor between the VNF's guest OS instance and the physical hardware as well as the impact of noisy-neighbors due to multi-tenancy and resource oversubscription, additional scheduling latency is introduced which can result in tardy triggering of isochronous interrupts and delays in scheduling resources (e.g. CPU) for VM instances. VM scheduling latency is the absolute value of the difference between when a guest operating system event should have executed and when the event actually executed. Figure 7 gives a Complementary Cumulative Distribution Function (CCDF) illustrating the actual notification latency for 1 millisecond timer events requested by a sample VNFC which illustrates this impairment. Tardy triggering of isochronous interrupts and VM scheduling often directly translates into jitter in response packets emitted by the VNF component executing in the impacted VM. The primary metrics are maximum scheduling latency and variance scheduling latency.

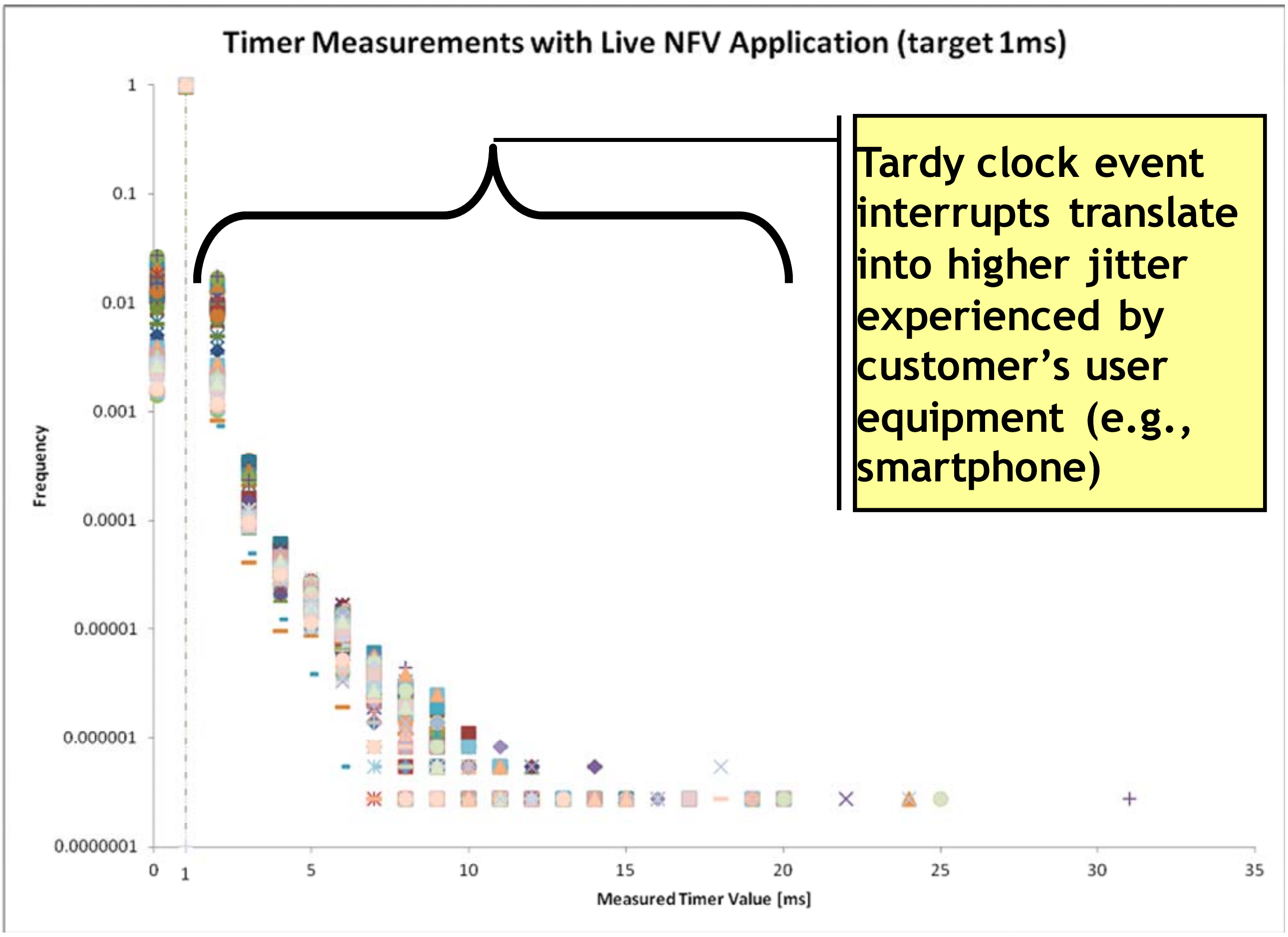

Figure 7: Sample Complementary Cumulative Distribution of VM Scheduling Latency of 1 msec Timer Events

5.4 VM Clock Error

VNF components often read a realtime clock to timestamp billing records, a provisioning or configuration change, creation or modification of a persistent data record, or record a fault or alarm event. While hardware realtime clock mechanisms generally have low and predictable clock drift and error, virtualisation and cloud computing can introduce material errors in the realtime clock value presented to VM instances, and that clock error can thus wander far more than with physical clocks. VM clock error characterizes the difference between the real time presented by the VM instance to the VNF and Universal Time.

6 Virtual Network Interface Service Quality Metrics

Virtual networking offered by NFV infrastructure communicates packets:

1) Between component instances within a single VNF.

2) Between a VNF's component instances and technology components used by that component, like database-as-a-service, load-balancing-as-a-service and storage-as-a-service.

3) Between the target VNF and other VNFs or traditional network elements.

The service boundary between VNF component instances and the NFV virtual network infrastructure is the VNFC's guest OS instance, meaning that the reference events for virtual network service quality metrics are:

1) when packets pass from sending VNFC instance to the underlying NFV infrastructure; and

2) when packets are passed from NFV infrastructure to receiving VNFC instance.

The virtual network (see ETSI GS NFV-INF 001 [1]) service quality metrics measured by the infrastructure network (e.g. NIC, vNIC) are:

1) **Packet loss** is the rate of packets that are either never delivered to the destination or delivered to the destination after the VNF's maximum acceptable packet delay.

2) **Packet delay** is the elapsed time between a packet being presented to the NFV virtual network from one VNFC's guest OS instance to that same packet being presented to the destination VNFC's guest OS instance. Packets that are delivered with more than the VNF's maximum acceptable packet delay are counted as packet loss events and excluded from packet delay measurements.

3) **Packet delay variance** (a.k.a. jitter) is the variance in packet delay.

4) **Network outage** characterizes virtual network impairments that persist for longer than the VNF's Maximum Acceptable Network Transient Time defined by the metric input parameters, or as an aspect of the SLA. Figure 8 illustrates how virtual network impairments that persist for no longer than the VNF's maximum acceptable network transient time are counted as packet loss episodes (undelivered packets or packets delivered too late, as defined earlier), and impairments that last longer are counted as network outage events. The key measurement parameters are rate of network outage events, duration of network outage events and impact extent of outage events. Network outage downtime and network availability can be derived from these network outage parameters.

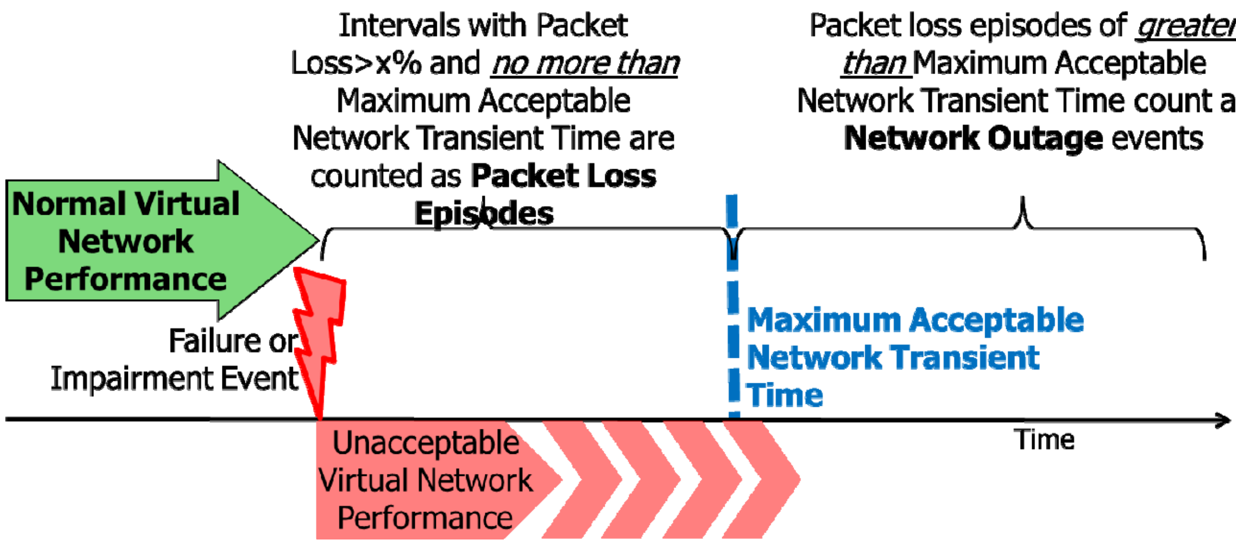

Figure 8: Virtual Network Packet Loss and Outage Downtime Metrics

7 Technology Component Service Quality Metrics

Platform-as-a-Service defined by [i.1] offers libraries, services and tools to NFVI consumers that consumers do not need to manage or control, such as database-as-a-service, load-balancing-as-a-service and storage-as-a-service. The Platform-as-a-Service provider who offers a technology component as a Technology-Component-as-a-Service to NFV consumers provides all ownership, operations, administration, maintenance and provisioning of the technology component which the NFV consumers would otherwise be responsible for if they obtained the underlying technology component as a virtual appliance or VNF directly from the software supplier.

VNF components interwork with technology components offered as-a-service via NFV virtual networking. The technology-component-as-a-service measurement demark is back-to-back with the virtual network measurement demark to assure that all loss/error, latency and unavailability impairment events are measured in one and only one service quality metric. Nominally the technology-component-as-a-service to NFV virtual networking demark will be at the guest OS for the technology component's frontend VNFC, but different back-to-back demark points can be used so long as no impairments in the service delivery path are omitted from both virtual network and technology component as-a-service quality metric coverage.

The core technology component as-a-service (TcaaS) quality metrics are:

1) **TcaaS Service Reliability** characterizes the portion of TcaaS operations or transactions that are correctly processed within the maximum acceptable TcaaS service latency.

2) **TcaaS Service Latency** characterizes the service latency of individual TcaaS operations or transactions. Transactions or operations which exceed the maximum acceptable latency are counted as TcaaS reliability impairments and are excluded from TcaaS latency metrics.

3) **TcaaS Service Outage** characterizes events when the TcaaS is unavailable to the VNF for longer than the maximum acceptable technology component transient time. As shown in figure 9, the service impact individual TcaaS service disruption events which persist for no more than the Maximum Acceptable Technology Component Transient Time (defined by an input parameter to the metric) are deemed failed operations and thus are charged as TcaaS service reliability impairments; disruptions for longer the Maximum Acceptable Technology Component Transient Time are charged as TcaaS Service Outage Downtime. The key measurement parameters are rate of TcaaS outage events, duration of TcaaS outage events and functionality and/or capacity impact of TcaaS outage events. TcaaS outage downtime and TcaaS availability can be derived from these TcaaS service outage parameters.

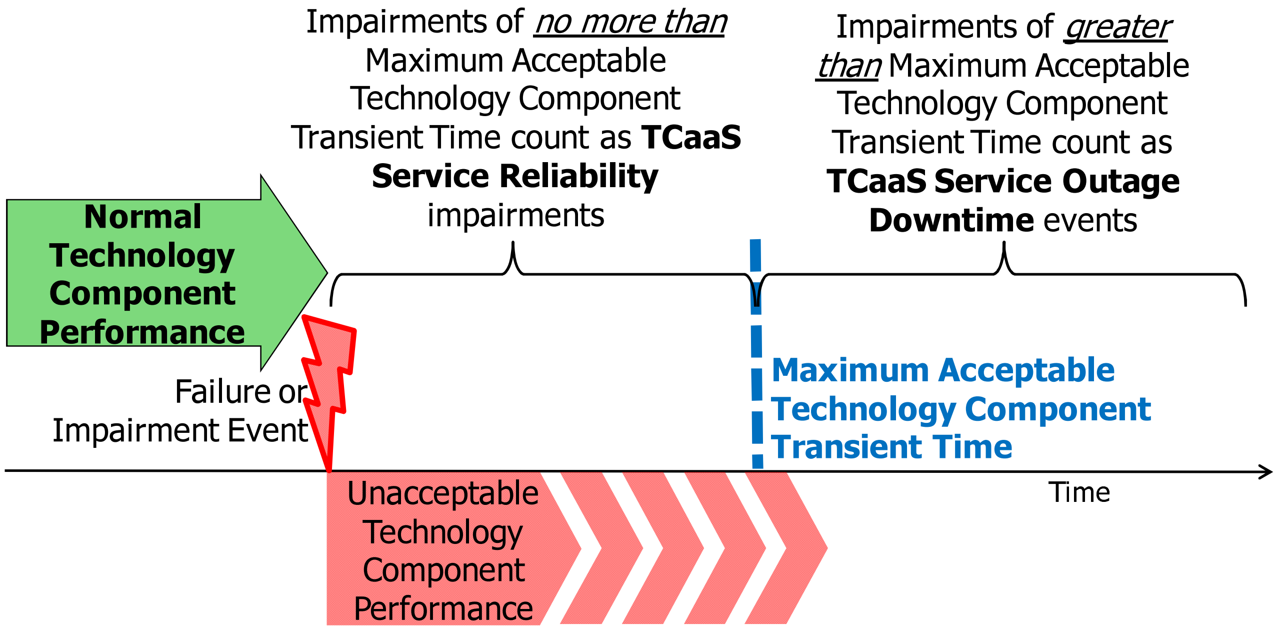

Figure 9: Technology Component as-a-Service Reliability and Outage Downtime Metrics

8 Orchestration Service Quality Metrics

8.1 Virtual Machine Orchestration

This clause describes the metrics on individual requests to establish virtual machines in more detail.

Figure 10 illustrates the orchestration service quality metrics in the context of the VM lifecycle.

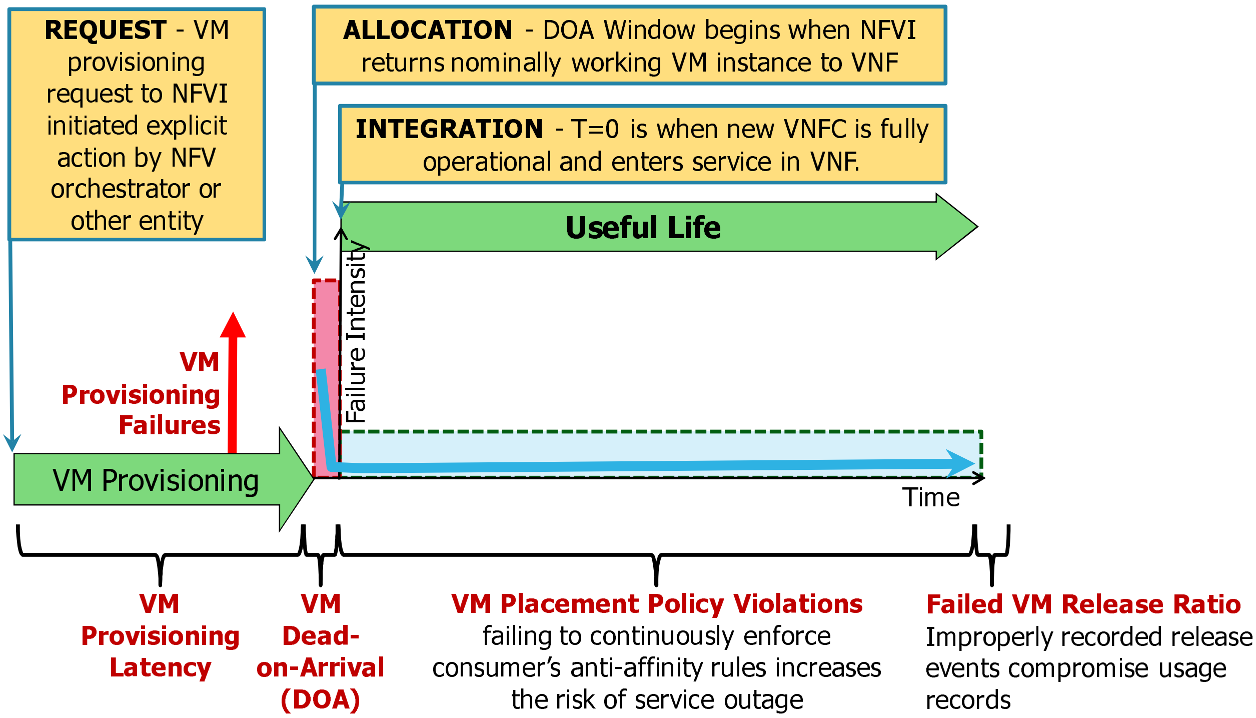

Figure 10: Orchestration Metrics on VM Lifecycle

1) **VM Provisioning Latency** is the elapsed time between a VM provisioning request being presented and the corresponding provisioning response being returned. Provisioning requests that take longer than the maximum acceptable VM provisioning latency are counted as VM provisioning reliability impairments and not counted in VM provisioning latency measurements.

2) **VM Provisioning Reliability** characterizes the ratio of VM provisioning requests that fail to complete successfully within the maximum acceptable VM provisioning time to total attempts. VM provisioning requests that fail due to syntax or semantics errors in the request may be excluded from the calculation. Note that only VM provisioning requests that return a failure indication are considered defective and counted against the VM provisioning reliability measurements; faulty VMs that are spuriously provided to VNFs and NFV consumers are counted via VM dead-on-arrival (DOA) metrics.

3) **VM Dead on Arrival (DOA)** characterizes the ratio (to total of nominally successful attempts) of VM instances that are delivered to the NFV consumer or VNF which purport to be fully operational but actually contain some latent configuration or other fault which prevents the VNFC hosted in the VM instance from entering user service (i.e. it never reaches the "T=0" event). Note that causes directly attributable to either the NFV consumer, VNF supplier or VNF itself are excluded from this measurement.

4) **VM Placement Policy Compliance** characterizes how rigorously the NFV consumer's antiaffinity rules for VM placement are continuously enforced. Rigorous enforcement of correct and complete antiaffinity rules assures that no single infrastructure failure event will overwhelm the VNF's high availability mechanisms to produce an unacceptably long user service disruption.

5) **Failed VM Release Ratio** characterizes how reliably the usage records for VM release are recorded. VM release events are charged as failures if the release event is not properly recorded in with a timestamp error of no more than the Maximum Acceptable Release Event Timing Error (for example, after a waiting timer on the release confirmation expires). This timestamp error is the difference between the UTC time when the VM release was requested and the recorded timestamp on the associated usage event record.

8.2 Virtual Network Orchestration

This clause describes the metrics on individual requests to establish virtual network connectivity in more detail. Discussion of extended metrics to assess the many requests needed to connect many VNFC or establish an entire VNF composed of multiple VM with VN connectivity follows.

1) **VN provisioning latency** metrics assess the time needed to successfully establish Infrastructure Network connectivity, which is the time from a request for connectivity to the time the success response is received at a measurement point.

2) **VN provisioning reliability** metrics track unsuccessful establishment attempts, or failure to establish Infrastructure Network connectivity when requested, determined from responses indicating failure or excess waiting time.

3) **VN diversity compliance** metrics characterize how accurately the NFV Infrastructure Network achieves the requested physical diversity, such as by achieving the desired number of physically diverse paths between endpoints.

Requests for Infrastructure Network Connectivity that involve multiple paths would be quantified by both the provisioning latency for the successfully established paths combined with the reliability of provisioning. For example, a metric on a request for 5 paths between endpoints is characterized by the pair (provisioning latency, provisioning reliability) or (10 ms, 80 %) when one path of five is unsuccessful). Note that the summary statistic on successful provisioning latency is to-be-determined (and may be out-of-scope for the SQM work item). The combination of the metrics on provisioning speed and reliability quantify partial success, which is preferred over a binary outcome.

Depending on how requests to instantiate the many components of VNFs are formulated, there appears to be a need to quantify the degree to which the instantiated resources satisfy the request(s) any time the VM or VN provisioning reliability is less than 100 %, or when VM placement policy and or VN diversity have not been fully satisfied.

9 Service Quality Metrics Use Case

This clause offers an illustrative vision of potential short term, medium term and long term uses of NFV service quality metrics. For simplicity, this use case considers only the 'maximum acceptable VM stall time' parameter of the VM Stall service quality metric. When considering actual use of service quality metrics one should keep in mind that different applications have different sensitivities to each of the service quality metrics. VM stall events directly impact isochronous delivery of bearer plane voice and video traffic, so unacceptably long VM stall events of VM instances hosting bearer plane VNFCs directly impact end user quality of service. In contrast, batch-oriented applications like bulk provisioning may not even detect VM stall events because occasional VM stall events will have negligible impact on provisioning throughput when measured as thousands of subscribers per hour. This example considers a hypothetical bearer plane VNF supporting Voice-over-LTE and an operations/business support system for that same VoLTE deployment.

9.1 Short Term Use of Service Quality Metrics

The initial use of objective and quantitative NFV service quality metrics is to facilitate rich conversations regarding service quality objectives between VNF suppliers, VNF consumers (e.g. organizations offering VoLTE service to end users via virtualised network elements) and NFV infrastructure service providers (e.g. organizations offering NFV orchestration, management and infrastructure services to VNF consumers). The VNF consumer delivering VoLTE service to end users will set a 'maximum acceptable VM stall time' objective that assures that end users experience acceptable bearer service quality even when occasional VM stall events occur. The VNF consumer can negotiate the service level objective for 'maximum acceptable VM stall time' with their NFV infrastructure service provider.

However, there would be no direct measurement of VM stall time, and approximate compliance with the service level objective would be inferred from other service quality metrics, such as bearer packet delay variation measured at intermediate points on the e2e service path to assist with impairment isolation.

Since the VoLTE OSS/BSS is relatively insensitive to VM stall events, the VNF consumer might not bother to set an explicit maximum acceptable VM stall time for that application, instead relying on the NFV service provider's default maximum acceptable VM stall time.

9.2 Medium Term Use of Service Quality Metrics

In the medium term, NFV hypervisors hosting VNFs will objectively measure VM stall events, and that data will be made available to VNF consumers via NFV management interfaces. When end users experience unacceptable voice quality, the VNF consumer can retrieve VM stall data (along with lost packet and other NFV service quality metrics data) to see if end user quality impairments are correlated with VM stall events (or with other measured NFV service quality impairments). After isolating the true root cause of the service impairment appropriate corrective actions can be taken.

9.3 Long Term Use of Service Quality Metrics

In the longer term NFV service providers might proactively manage their infrastructure and workload based on service level objectives/agreements (SLOs/SLAs) based on NFV service quality metrics. For example, the VNF consumer might set the maximum acceptable VM stall time for VoLTE bearer plane VNFs to 10 milliseconds and set the maximum acceptable VM stall time for VoLTE operations/business support systems to 2 000 milliseconds. The NFV service provider would then carefully place the bearer plane VNFs into positions so they should rarely need to be moved (i.e. live migrated), and the hypervisor would be configured to assure that those VNFs experienced minimal VM/resource stall/disruption; the NFV service provider might charge the VNF consumer higher prices for these premium 'no-stall' VMs. In contrast, the VoLTE OSS/BSS VNFs can easily tolerate moderate VM stall events, so the NFV service provider has far greater freedom to live migrate those VMs from one host to another to better manage their data center resources; the NFV service provider might charge the VNF consumer lower prices for these easier-to-manage VMs.

10 Recommendations

This clause describes the recommendations meant to ensure the adequate understanding and treatment for Service Quality Metrics in the NFV framework. Some aspects concern follow-on activities in different Standards Development Organizations, others deal with more practical considerations of how metrics are produced and used. IThe current GS scope excluded work in some areas that may need to be addressed in external foras.

10.1 Measurement of Service Quality Metrics

This clause expresses two requirements pertaining to SQM:

- Requirement SQM1: NFV components which implement and deliver the services that are covered by these metrics to VNFs shall support the capability to measure the applicable metrics.
- Requirement SQM2: measurement control and reporting shall be performed through a management system, including the ability to start/stop measurement activity.

EXAMPLE: The NFV Infrastructure elements that implement the [Vn-Nf]/VM should measure the Virtual Machine metrics and the elements that implement the [Vn-Nf]/N should measure the Virtual Network metrics. VNFs should also measure the metrics that are most influential on end user service quality. Both NFV service providers and consumers should be prepared to audit service provider measurements against consumer measurements to reconcile any inconsistencies.

10.2 Service Quality Metrics in SLAs

The service quality metrics are appropriate to use when defining both service level objectives (SLOs) and service level agreements (SLAs) for NFV infrastructure, management and orchestration services delivered to VNF instances. Service quality metric objectives or agreements are likely to influence operational aspects of NFV infrastructure, management and orchestration to minimize the risk of NFV performance failing to meet targets. All references to SLAs and SLOs throughout NFV documents should consider appropriate service quality metrics as a use case.

10.3 Detailed Metric Definitions

This document describes a set of metrics in sufficient detail to communicate the quantities assessed and illustrate their importance and value. However, more details are necessary to establish unambiguous metrics definitions, such that independent implementations of the metrics would produce exactly the same measured values under identical conditions. For example, IETF first prepared a framework in IETF RFC 2330 [i.2] describing the common concepts needed for most packet transfer measurement. ETSI GS NFV-PER 001 [i.6] describes the development guidelines metrics more generally. But there are many areas described above where both the foundational framework and metrics definitions will require new development, and it is recommended to identify a suitable standards venue and pursue new work there.

10.4 Methods of Measurement

Implementing a measuring system to assess a standard metric involves embracing practical realities, and also requires agreement on approximations, inferences, and sources of error incurred in the standard methods of measurement. Thus it is important to couple each standardized performance metric with one or more methods of measurement applicable to a range of measurement circumstances.

10.5 Reporting Statistics and Results Processing

Reporting repeating results from a measuring system requires agreement on the summary statistics to apply, as different systems reporting in any way they choose represents a challenge to any attempts to compare results and effectively obscures any benefit from standard metrics and methods. In long-term collection, exceptional conditions are inevitable and may corrupt the summary if left unmitigated. However, it is necessary to codify the counting or exclusion rules to remove the impaired or questionable results from the summary to achieve consistency.

10.6 Characterization Plans

Recommendations on how to thoroughly assess a complete system - what to measure, where, and how often - are usually valuable as a starting point in new installations. Thus the development of at least one representative characterization plan for the key NFV use cases is recommended. A complete plan will identify measurement points in the overall NFV architecture, as well as supplying numerical values for the metric parameters (such as thresholds of unacceptable waiting time).

Annex A (informative): Example NFVI-VIM Interactions related to Service Quality Metrics

A.1 Introduction

This annex describes some of the key operations where the VIM and domains of the NFVI interact, and examines the events and notifications associated with the different operations communicated across the [Nf-Vi]/<C,N,H> reference points related to the key service quality metrics described in the body above. The annex also makes these examples available as a reference to other documents.

A.2 Resource Establishment (VM)

In this clause, focuses on the VIM Request-for/Response-to establishment of NFVI components, specifically the Virtual Machine (VM). The performance metrics listed in the present document for instantiation of virtual resources is considered. For VM, there is (part of table 1):

Table A.1

Service Metric Category	Speed	Accuracy	Reliability
Orchestration (e.g. Resource Allocation, Configuration and Setup)	VM Provisioning Latency	VM Placement Policy Compliance	VM Provisioning Reliability VM Dead-on-Arrival (DOA) Ratio

A.2.1 VM Provisioning Latency

This metric intends to track the time needed for a VIM to successfully instantiate a VM in the NFVI, which is a simple Request-Response exchange:

- Request: Begin establishment of a valid VM provisioning operation; a coherent computing resource with specified pool id, memory, storage, vNIC, etc.
- Return: Hypervisor Event indicating **successful** establishment of the requested VM.

Events and VM Management Information are available. For example, a hypervisor could indicate that it has "stood-up" a VM through the notification that a VM is Running, delivered as a message from the Hypervisor using the "Virtual Machine Monitoring MIB":

```
The OID tree structure of the MIB module is shown below.

 --vmMIB (1.3.6.1.2.1.yyy)
   +--vmNotifications(0)
   |  +--vmRunning(1) [vmName, vmUUID, vmOperState]
   …
```

For measurement purposes, the VIM could store the timestamp of the Request (t1) and use the corresponding Return arrival timestamp (t2) of the response as the basis for latency calculation (t2 - t1).

The example notification above does not include a source timestamp, thus it should be read and time stamped by the VIM, as soon as possible on reception.

Table A.2

Operations	Description	VIM Input	VIM Output	Notes
Request	This operation Instantiates a VM	Request from Orchestration	Request message	Event "T1"
Return	Immediate Notification of VM existence and status	"running" Notification from Hypervisor	Notify Orchestration (also Immediate)	Event "T2" (time ≤ Tmax)
Measurement	Calculate service quality metric, if successful		VM Provisioning Latency metric	T2 - T1

A.2.2 VM Provisioning Reliability

This metric intends to track the inability of the NFVI to instantiate the requested VM. The Request-Response exchange would proceed as follows:

- request: Begin establishment of a VM; a coherent computing resource with specified pool id, memory, storage, vNIC, etc.
- return: Hypervisor Event indicating **rejection** of the requested VM; or
- time-out: No response to the Request within **Tmax**.

The example/candidate notifications for **rejection** of VM Instantiation delivered as a message passed from the Hypervisor using "Virtual Machine Monitoring MIB" include the following VirtualMachineOperState (s):

```
unknown(1)     The operational state of the virtual
               machine is unknown, e.g. because the
               implementation failed to obtain the state
               from the hypervisor.

Other(2)       The operational state of the virtual
               machine indicating that an operational
               state is obtained from the hypervisor but
               it is not a state defined in this MIB
               module.
```

For measurement purposes, the VIM could store a count of rejections and timeouts, and provide current counts (or ratio of reject+timeout counts to total attempts in the measurement interval).

Table A.3

Operations	Description	VIM Input	VIM Output	Notes
Request	This operation Instantiates a VM	Request from Orchestration	Request message	total_attempts++
Return	Immediate Notification of VM status	"unknown/other" Notification from Hypervisor	Notify Orchestration (also Immediate)	Reject++ (time ≤ Tmax)
Request Time-out	Waiting time Tmax exceeded	Time-out interrupt (internal)	Notify Orchestration (Immediate)	Timeout++ (even if eventually successful)
Calculate	Calculate service quality metric(s)		VM Provisioning Reliability metric	Current counts or ratio of (Rej+Timeout)/total

A.2.3 VM DOA Ratio

This metric intends to track the liveliness of the requested VM. The Request-Response exchange would proceed as follows:

- request: Begin establishment of a VM; a coherent computing resource with specified pool id, memory, storage, vNIC, etc. - see section 8.3 of [i.12];

- return: Hypervisor Event indicating **successful** establishment of the requested Vm; and
- the VM is found to non-working. For example the VirtualMachineOperState indicates, either in the Return or immediately thereafter, a non-running and non-normal operation state.

The candidate notifications for DOA (Dead on Arrival) VM Instantiation delivered as a trap from the "Virtual Machine Monitoring MIB" include the following VirtualMachineOperState (s):

```
blocked(5)     The operational state of the virtual
               machine indicating the execution of the
               virtual machine is currently blocked,
               e.g. waiting for some action of the
               hypervisor to finish.  This is a
               transient state from/to other states.
Crashed(13)    The operational state of the virtual
               machine indicating the virtual machine
               has crashed.
```

For measurement purposes, the VIM could store a count of DOA notifications, and provide current counts (or ratio of DOA counts to total attempts in the measurement interval).

Table A.4

Operations	Description	VIM Input	VIM Output	Notes
Request	This operation Instantiates a VM	Request from Orchestration	Request message	total_attempts++
Return	Immediate Notification of VM existence and status	"blocked/crashed" Notification from Hypervisor	Notify Orchestration (also Immediate)	DOA++
Calculate	Calculate service quality metric		VM DOA Ratio metric	DOA / total

A.2.4 VM Placement Policy Compliance

This metric intends to track enforcement of the anti-affinity parameter in a provisioning request. The information to determine VM instance affinity (with other VMs, for example) *could* be supplied as a hierarchy of infrastructure component identifiers supporting the specific instance:

Table A.5

VM	Hypervisor	CPU/Core	NIC	Storage
a	b	c-d	e	F

NOTE 1: The current VM Service Description does not mention anti-affinity explicitly, though it may be an aspect of the "resiliency requirements" Request Parameter and "resiliency" Return Parameter.

Since the VIM has HW&SW inventory and other information, it appears that responsibility for attaining the desired degree of anti-affinity falls to the VIM (and not Infrastructure).

NOTE 2: The placement policy section in very preliminary and requires more work.

A.3 Resource Establishment (Infrastructure Network)

In this clause focuses on the VIM Request-for/Response-to establishment of NFVI components, specifically the Virtual Network (VN) within the Infrastructure Network.

NOTE: The need for establishment and reservation of individual links in the Infrastructure Network is under discussion. ETSI GS NFV-MAN 001 [i.4] puts the role of resource allocation on the VIM, and the informative Annex B talks about reserving resources as part of a feasibility step, prior to instantiation, which could apply to the virtual network resources as well as others.

For VN Establishment, a new row of performance metrics listed in the present document for instantiation of virtual resources (which is part of table 1) is required:

Table A.6

Service Metric Category	Speed	Accuracy	Reliability
Establishment	VN Provisioning Latency	VN Diversity Compliance	VN Provisioning Reliability

A.3.1 VN Provisioning Latency

This metric intends to track the time needed for a VIM to successfully instantiate a VN in the NFVI, which is a simple Request-Response exchange :

- Request: Begin establishment of a VN; a logical connectivity between endpoints with specified pool id, redundancy, bandwidth, latency, etc. - see table 2 section 8.3 of [i.12].
- Return: Infrastructure Control Event indicating **successful** establishment of the requested VN.

Table A.7

Operations	Description	VIM Input	VIM Output	Notes
Request	This operation Instantiates a VN	Request from Orchestration	Request message	Event "T1"
Return	Immediate Notification of VN existence and status	"success" Notification from NFVI	Notify Orchestration (also Immediate)	Event "T2" (time<=Tmax)
Calculate	Calculate and service quality metric, if successful		VN Provisioning Latency metric	T2 - T1

A.3.2 VN Provisioning Reliability

This metric intends to track the inability of the NFVI to instantiate the requested VN. The Request-Response exchange would proceed as follows:

- request: Begin establishment of a valid VN provisioning operation; a logical connectivity between endpoints with specified pool id, redundancy, bandwidth, latency, etc. - see section 8.3 of [i.12];
- return: Infrastructure Network Control Event indicating **rejection** of the requested VN; or
- time-out: No response to the Request within **Tmax**.

Table A.8

Operations	Description	VIM Input	VIM Output	Notes
Request	This operation Instantiates a VN	Request from Orchestration	Request message	total_attempts++
Return	Immediate Notification of VN status	"reject" Notification from NFVI	Notify Orchestration (also Immediate)	Reject++ (time<=Tmax)
Request Time-out	Waiting time Tmax exceeded	Time-out interrupt (internal)	Notify Orchestration (Immediate)	Timeout++ (even if eventually successful)
Calculate	Calculate service quality metric(s)		VN Provisioning Reliability metric	Current counts or ratio of (Rej+Timeout)/total

Annex B (informative): Authors & contributors

The following people have contributed to the present document:

Rapporteur:

Dr. Julien Maisonneuve, Alcatel-Lucent

Other contributors:

Mr. Eric Bauer, Alcatel-Lucent
Mr. Al Morton, AT&T Labs

Annex C (informative): Bibliography

- VIM Southbound Interfaces - Michael Brenner (ALU)
  http://docbox.etsi.org/ISG/NFV/MAN/05-CONTRIBUTIONS/2014//NFVMAN(14)000159_VIM_Southbound_interfaces_placeholder.doc

- Current Data supplied to the VIM from INF - Valerie Young (Intel)
  http://docbox.etsi.org/ISG/NFV/MAN/05-CONTRIBUTIONS/2014/NFVMAN%2814%29000150_Current_Data_supplied_to_the_VIM_from_INF.docx

- Inclusion of VM and VN service Descriptions in Domain Documents- Andy Reid (BT)
  http://docbox.etsi.org/ISG/NFV/MAN/05-CONTRIBUTIONS/2014//NFVMAN(14)000147r1_Inclusion_of_VM_and_VN_service_Desciptions_in_Domain_Doucmen.docx

- "Management Information Base for Virtual Machines Controlled by a Hypervisor" Asai, H. et al., IETF Stream Internet Draft intended for Standards Track, February 10, 2014, work-in-progress, http://tools.ietf.org/html/draft-ietf-opsawg-vmm-mib-00

History

Document history		
V1.1.1	December 2014	Publication