QuEST Forum

TL 9000
Quality Management System
Committee Final 1.0

NFV Workload Efficiency Whitepaper

For further information,
see the QuEST Forum Web page at:
http://www.questforum.org/

NFV Strategic Initiative Team

Adtran

Ed Bryan

AT&T

Cameron Shearon
Beth Ford
Brad Ruark
Ray Fannon

Avacend

Kanchana Raman (Board Sponsor)

CenturyLink

Michael Fargano
Bill O'brien
Dennis Petrie
Joseph Diel
Michael Bugenhagen

Cisco

Sheronda Jeffries

DISH Networks

Bill Parent

Ericsson

Maria Eriksson

Fujitsu

Ashok Dandekar
Zigi Putnins
Hatsumi Lino
Katsumi Fukumitsu
Masahiro Simbashi
Motoyoshi Sekiya
Nahohito Taga

Genband

Paul Smith

Hitachi

Paul Neighbour
David Blackwelder
Anand Krishnaswamy

Huawei

Don Topper

Infinera

Stephen Choy

Juniper Networks

Pasvorn Boonmark

KCI Telecommunications

Charlie Millard
Sue Breitsprecher

Nokia

Eric Bauer, Author
Ben Jernigan
John Wronka

Oracle

Leslie Smith
Madonna Comeau
Sarah Cornel

Overture Networks

Prayson Pate
Barry Shapiro

QuEST Forum

Tom Yohe
Ken Koffman

University of Texas at Dallas

Ron Bose

Verizon Wireless

Michael Beard
Nathan Jeffrey
Abhitabh Kushwaha
Shane Ronan
Doug Grupe

ZTE Corporation

Dan Chen

Abstract

A standard goal of cloud is:

> *From the customers' perspective,* ***cloud computing offers the users value by enabling a switch from a low efficiency and asset utilization business model to a high efficiency one***; (ISO/IEC, 2014-10-15)

Fundamentally, workload efficiency is the ratio of application output delivered to cloud service users divided by the resource inputs to that application service, like cloud infrastructure, management, orchestration and functional component provided as-a-service. This whitepaper frames NFV workload efficiency quality attributes suitable for the following use cases:

- Estimate Likely Operating Expenses of an NFV-based Service or Solution (section 2.1)
- Workload Efficiency Measurements in Supplier Selection (section 2.2)
- Workload Efficiency Measurements in Ongoing Operations (section 2.3)
- Workload Efficiency Measurements in Performance SLAs (section 2.4)
- Benchmark Workload Efficiency Performance (section 2.5)

Section 3 considers *Efficiency as a Quality* and section 4 reviews *Why Is Efficiency More Complicated for NFV?* Section 5 discusses *Modeling VNF Resource Usage* and section 6 *Quantitatively Measuring Efficiency Qualities* argues that NFV workload efficiency measurements should be based on the ISO/IEC 15939 System and Software Measurement Process standard. The paper goes on to consider the following NFV workload efficiency quality attributes:

- ✓ **VNF Resource Efficiency** (section 7.1) – characterizes the ratio of useful service output delivered by a VNF as a function of virtualized infrastructure resource consumption across the Vn-Nf reference point.

- ✓ **VNF Elasticity Efficiency** (section 7.2) – characterizes how close a VNF's online capacity can come to a cloud service customer's online application capacity target.

- ✓ **VNF Lifecycle Management Automation Efficiency** (section 7.3) – characterizes the efficiency of VNF automated lifecycle management actions.

- ✓ **Energy Efficiency of NFV Infrastructure** (section 8.1) – characterizes the ratio of virtualized infrastructure services delivered to VNFs across the Vn-Nf reference point to energy input.

- ✓ **Virtualization Efficiency of NFV Infrastructure** (section 8.2) -

- ✓ **NFV Management and Orchestration Efficiency** (section 9)

Section 10 considers *Evolving Energy Efficiency Measurement*, section 11 offers *Proposed Requirements for NFV Efficiency Measurements* and section 12 gives *Recommendations*.

Contents

1 Efficiency Basics

We begin with several fundamental principles:

- Definition of Efficiency (section 1.1)
- Efficiency as a Ratio (section 1.2)
- Elasticity, Capacity and Utilization (section 1.3)
- Why Care About Efficiency? (section 1.4)

1.1 Definition of Efficiency

Merriam-Webster offers the following definitions of ***efficiency***[1]:

> *2a. efficient operation*
> *b(1): **effective operation as measured by a comparison of production with cost (as in energy, time and money)***
> *b(2): the ratio of the useful energy delivered by a dynamic system to the energy supplied to it.*

Automobile fuel efficiency is a well understood efficiency measurement. Fuel efficiency in the United States is measured as the distance traveled (i.e., the useful output of an automobile) by the fuel consumed (i.e., the resource input) traveling that distance. Some countries express fuel efficiency as liters of fuel consumed traveling 100km (i.e., resource input divided by useful output), but the notion of efficiency as the relationship between useful output and resource input still holds.

ITIL offers a similar definition of ***efficiency***:

> *A measure of whether the right amount of resource has been used to deliver a process, service or activity. An efficient process achieves its objectives with the minimum amount of time, money, people or other resources.* (Axelos Limited, 2011)

Table 1-1 contrasts the ITIL definitions efficiency (e.g., relationship of service output to resource input) with effectiveness (i.e., whether output/result achieves the objectives). Effectiveness measurements are not considered in this whitepaper.

Table 1-1 Contrasting ITIL Efficiency and Effectiveness (from (Axelos Limited, 2011)

Concept	Efficiency	Effectiveness
Definition	***A measure of*** *whether the right amount of* ***resources*** *have been* ***used to deliver a process, service or activity***	***A measure of whether the objectives*** *of a process, service or activity* ***have been achieved***
Example	*An efficient process achieves its objectives with the minimum amount of time, money, people or other resources*	*An effective process or activity is one that achieves its agreed objectives.*

TM Forum offers numerous operational efficiency measurements (enumerated in section 14 *Annex A - TM Forum Operational Efficiency Metrics*), however none of these are directly applicable to NFV workload efficiency. While ITIL offers numerous key performance indicators (see http://wiki.en.it-processmaps.com/index.php/ITIL_Key_Performance_Indicators), ITIL does not offer any KPIs explicitly related to NFV workload efficiency.

1.2 Efficiency as a Ratio

Figure 1-1 offers a canonical model of an IT service component like a VNF. The VNF --- technically an ITIL ***configuration item***[2] --- consumes infrastructure resources as input (e.g., compute, memory, networking, storage) and delivers some IT service as output.

[1] http://www.merriam-webster.com/dictionary/efficiency retrieved 5/16/16.
[2] ***Configuration Item*** is defined by ITIL as:

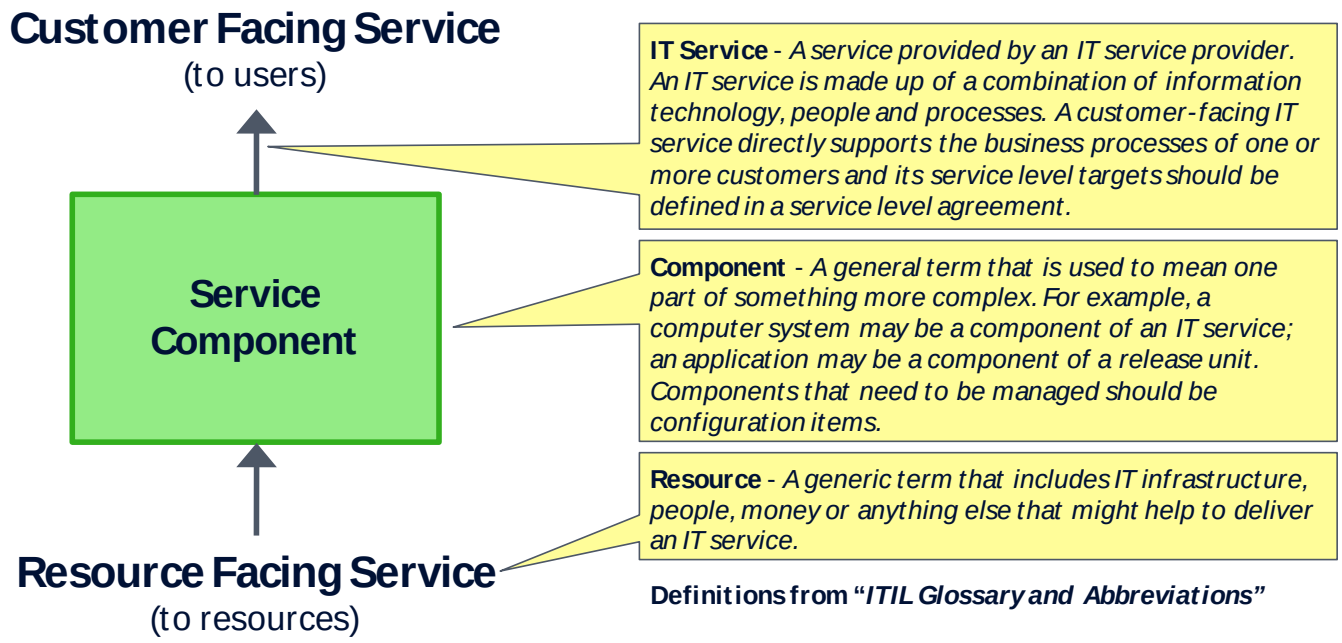

Figure 1-1 Canonical Service Component Model

Figure 1-2 visualizes efficiency of a target service component as the quantity of customer facing service output divided by the quantity of resource facing service input. Reducing the quantity of resource facing service input consumed to produce a fixed output by a service component (or producing more output with the same resource input) improves its efficiency.

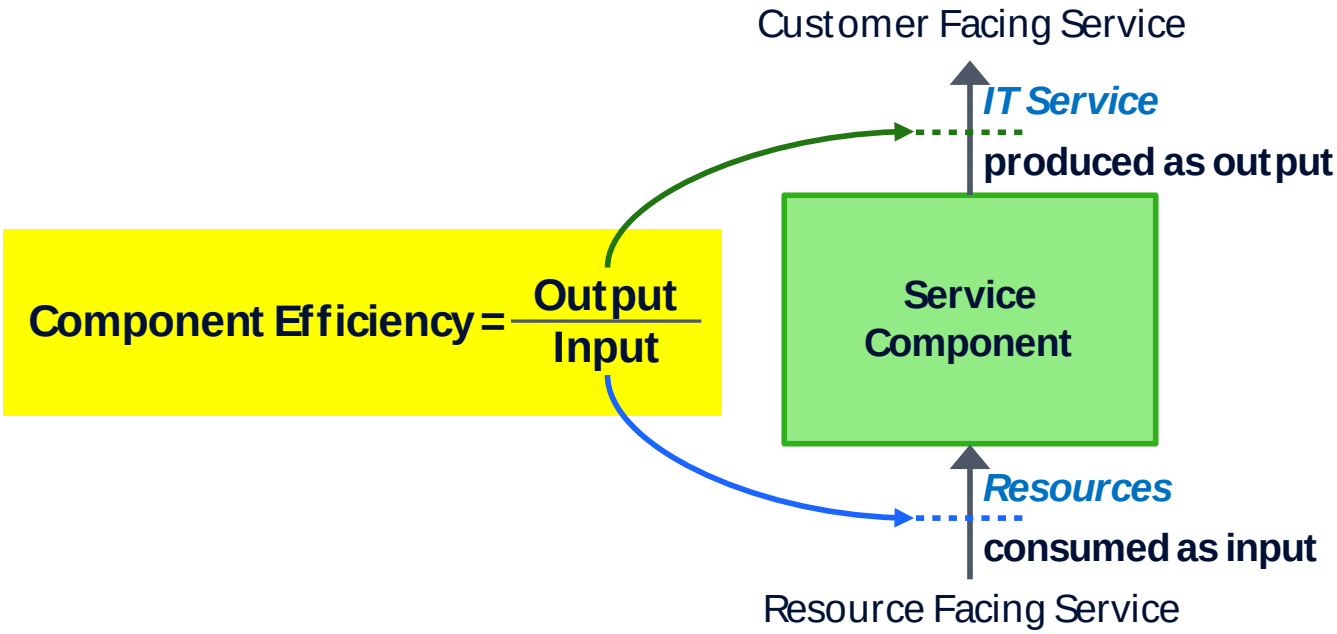

Figure 1-2 Component Efficiency as a Ratio

Any component or other service asset that needs to be managed in order to deliver an IT service. Information about each configuration item is recorded in a configuration record within the configuration management system and is maintained throughout its lifecycle by service asset and configuration management. Configuration items are under the control of change management. They typically include IT services, hardware, software, buildings, people and formal documentation such as process documentation and service level agreements. (Axelos Limited, 2011)

Figure 1-3 applies this common meaning of VNF resource efficiency to the canonical NFV architecture: fundamentally, workload efficiency is the ratio of application output delivered to cloud service users divided by the resource inputs to that VNF (e.g., across the Vn-Nf NFV reference point), like cloud infrastructure, management, orchestration and functional component provided as-a-service. To minimize operating expense, cloud service customers want to engage the least cost --- presumably the smallest --- fleet of cloud resources to serve a particular user workload.

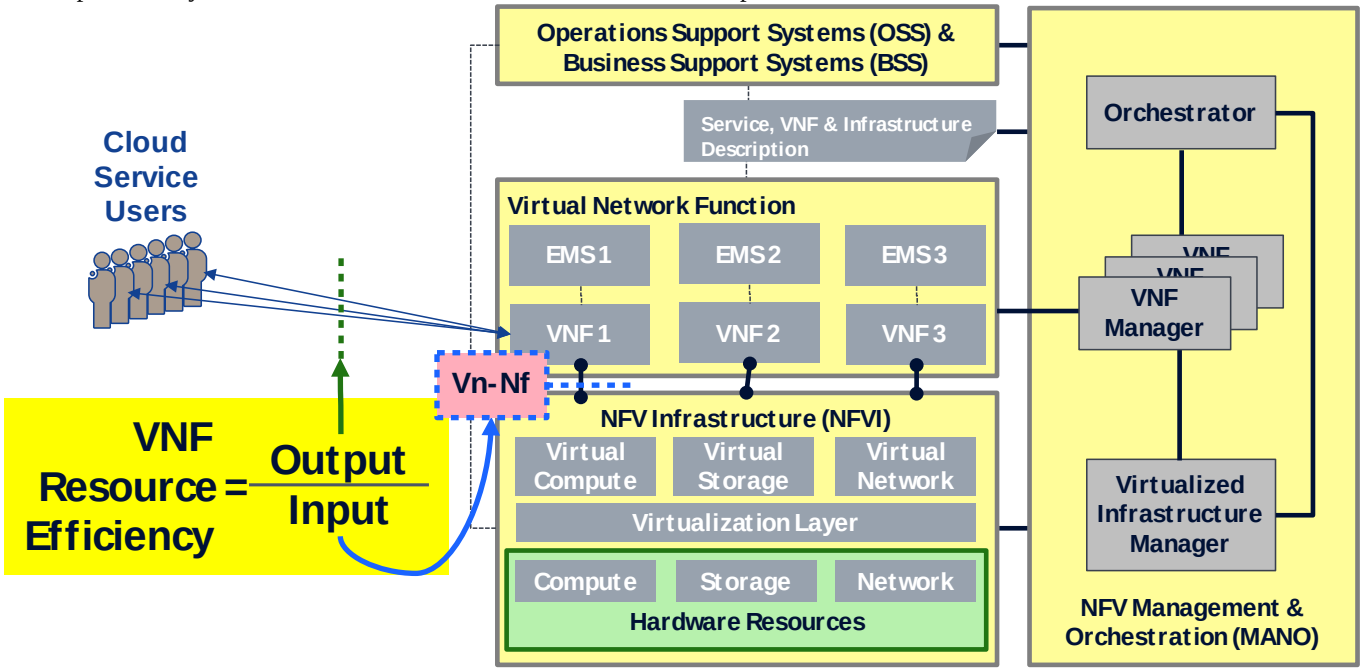

Figure 1-3 Steady State VNF Workload Efficiency of Cloud-Based Applications

1.3 Elasticity, Capacity and Utilization

ITIL defines ***capacity*** as:

> *The maximum throughput that a configuration item or IT service can deliver. For some types of CI, capacity may be the size or volume – for example, a disk drive*

In the context of capacity ratings, *maximum throughput* is generally understood to mean *maximum* ***steady state*** *throughput* ***with acceptable service quality, especially acceptable reliability and performance***.

ISO/IEC/IEEE 24765 defines ***configuration*** as:

> *the arrangement of a computer system or component as defined by the number, nature, and interconnections of its constituent parts*

A VNF's configuration thus includes the number, nature and interconnections of the resources assigned to the VNF; that configuration determines the maximum throughput that the VNF can deliver.

ISO/IEC 17788 stipulates:

> ***Rapid elasticity and scalability****: A feature where physical or virtual resources can be rapidly and elastically adjusted, in some cases automatically, to quickly increase or decrease resources. For the cloud service customer, the physical or virtual resources available for provisioning often appear to be unlimited and can be purchased in any quantity at any time automatically, subject to constraints of service agreements…*

Rapid elasticity and scalability enables one to reconfigure the maximum throughput (i.e., capacity) that a VNF can deliver rapidly to respond to user demand and business needs. However, at any instant in time a particular VNF has a deterministic resource configuration which establishes the maximum throughput that the VNF can deliver at that instant (i.e., its capacity).

ISO/IEC/IEEE 24765 defines ***utilization*** as:

> *in computer performance evaluation, a ratio representing the amount of time a system or component is busy divided by the time it is available*

Operationally, VNF capacity is about the resources that have allocated and configured for that VNF while utilization is about the portion of the configured resource allocation that is busy (i.e., utilized). Consider two scenarios:

- If *utilization* of a VNF's current online *capacity* is too **high**, then orchestration mechanisms will leverage rapid elasticity to allocate and add additional resources to change the VNF's configuration and increase its maximum throughput / *capacity*.

- If *utilization* of a VNF's current online *capacity* is too **low**, then orchestration mechanisms will reconfigure the VNF to a smaller configuration with a smaller maximum throughput / *capacity* and return the excess resources to the cloud service provider.

1.4 Why Care About Efficiency?

Inevitably, some resources are scarce/finite and precious/costly. The organization that owns and operates a service component naturally wishes to limit consumption of the resources that are scarce/finite and/or precious/costly to them, including: time; human labor; and infrastructure services like compute, memory, storage and networking. For example, cloud service providers (CSPs) may have finite physical space, electrical power and cooling capacity in a data center, so they will carefully select equipment to efficiently consume those finite and/or costly resources. Cloud service customers ultimately must pay for the virtualized compute, memory, storage and networking offered by CSPs that they consume, so higher efficiency VNFs should consume less CSP resources as input to produce a unit of application service output. Reducing the intensity of input resource consumption for a given unit of service output ultimately improves the operational efficiency, and presumably business results, of any organization.

Any party that pays for resource inputs should benefit from higher efficiency as that lowers the quantity of resource input they consume and ultimately pay for. Thus, cloud service customers are likely to be concerned with VNF efficiency while infrastructure-as-a-service cloud service providers will be concerned with NFV infrastructure efficiency.

2 Target Use Cases

NFV efficiency measurements should be useful in the following scenarios:

- Estimate Likely Operating Expenses of an NFV-based Service or Solution (section 2.1)
- Workload Efficiency Measurements in Supplier Selection (section 2.2)
- Workload Efficiency Measurements in Ongoing Operations (section 2.3)
- Workload Efficiency Measurements in Performance SLAs (section 2.4)
- Benchmark Workload Efficiency Performance (section 2.5)

2.1 Estimate Likely Operating Expenses of an NFV-based Service or Solution

The resource costs that a cloud service provider or cloud service customer must pay to operate a solution is impacted by both the level of resources consumed serving workload as well as the price of those consumed resources. CSC and CSP organizations can estimate their operating expenses for offering a service by combining their estimated service workload with resource efficiency measurements of relevant service components (in scope of this paper) as well as the price they expect to pay for those resources (beyond the scope of this paper).

2.2 Workload Efficiency Measurements in Supplier Selection

Customers can use standard efficiency rating values from suppliers when evaluating competing product offerings. For example, VNF resource efficiency ratings are likely to be considered in CSCs' purchasing decisions and NFV infrastructure efficiency ratings are likely to be considered in CSPs' purchasing decisions.

2.3 Workload Efficiency Measurements in Ongoing Operations

As efficiency directly impacts a customer's operating expenses, and operating expenses directly impacts the customer's business performance, customers are likely to periodically (e.g., quarterly) compute actual efficiency. Customers may initiate corrective actions if efficiency fails to meet their expectations and/or drive continuous improvement activities to further improve efficiency.

2.4 Workload Efficiency Measurements in Performance SLAs

Some customers may wish to share the efficiency (i.e., resource consumption) risk with their suppliers via performance SLAs with some remedy if actual efficiency is materially worse than expected efficiency.

2.5 Benchmark Workload Efficiency Performance

Objective and quantitative measurements of NFV efficiency can be published via QuEST Forum mechanisms to make (anonymous) best-in-class, worst-in-class and industry average data on actual NFV efficiency performance available to QuEST Forum members to drive efficiency improvement across the industry.

3 Efficiency as a Quality

ISO/IEC 25000 "*Software product Quality Requirements and Evaluation (SQuaRE)*" family provides the most authoritative, model for non-functional qualities of software products and software-based services, including efficiency. ISO/IEC 25000 defines ***quality model*** as "*defined set of characteristics, and of relationships between them, which provides a framework for specifying quality requirements and evaluating quality*" (ISO/IEC, 2005-08-01). Figure 3-1 illustrates the general ISO/IEC 25000 quality model:

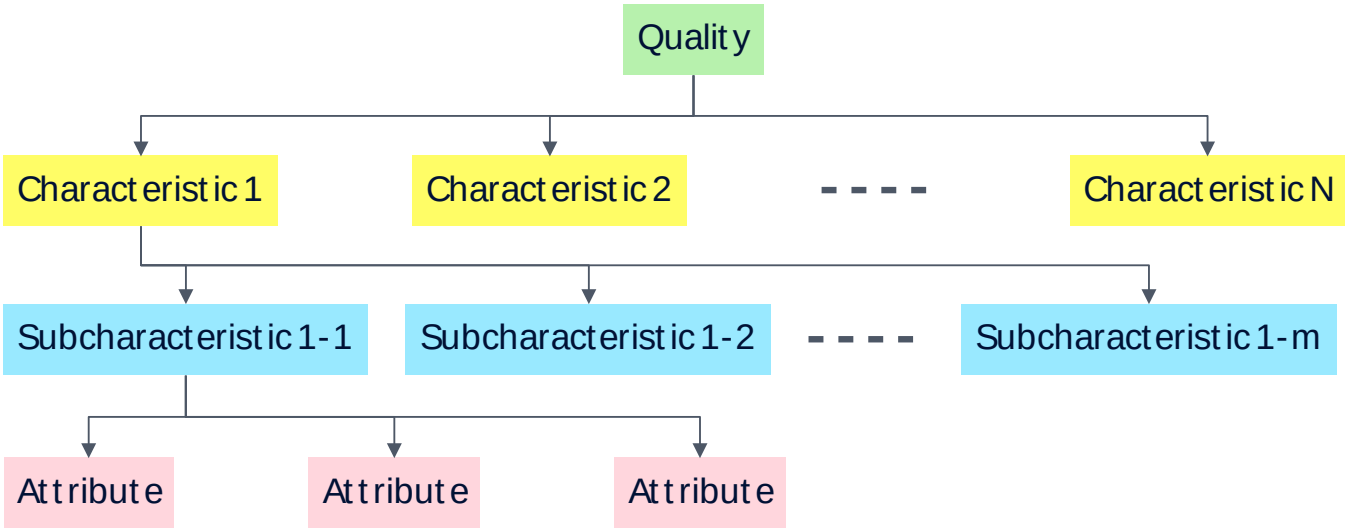

Figure 3-1 ISO/IEC 25000 Quality Model (from Figure 4 of (ISO/IEC, 2005-08-01))

- **Quality**, such as quality in use (ISO/IEC 25010 clause 4.1 (ISO/IEC, 2011-03-01)), product quality (ISO/IEC 25010 clause 4.2 (ISO/IEC, 2011-03-01)) or data quality (ISO/IEC 25012 (ISO/IEC, 2008-12-15))

- **Characteristics**, such as effectiveness, efficiency and satisfaction as characteristics of quality in use

- **Subcharacteristics**, such as usefulness, trust, pleasure and comfort as subcharacteristics of the satisfaction characteristic of quality in use

- **Attribute**, defined as "*inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means*" (ISO/IEC, 2005-08-01)

The ISO/IEC 25000 family of standard includes three types of quality that contain efficiency characteristics:

- *Efficiency* as a ***quality-in-use*** attribute from ISO/IEC 25010 clause 4.1; see Figure 3-2
- *Performance efficiency* as a ***product quality*** attribute from ISO/IEC 25010 clause 4.2; see Figure 3-3
- *Efficiency* as a ***data quality*** from ISO/IEC 25012; see Figure 3-4

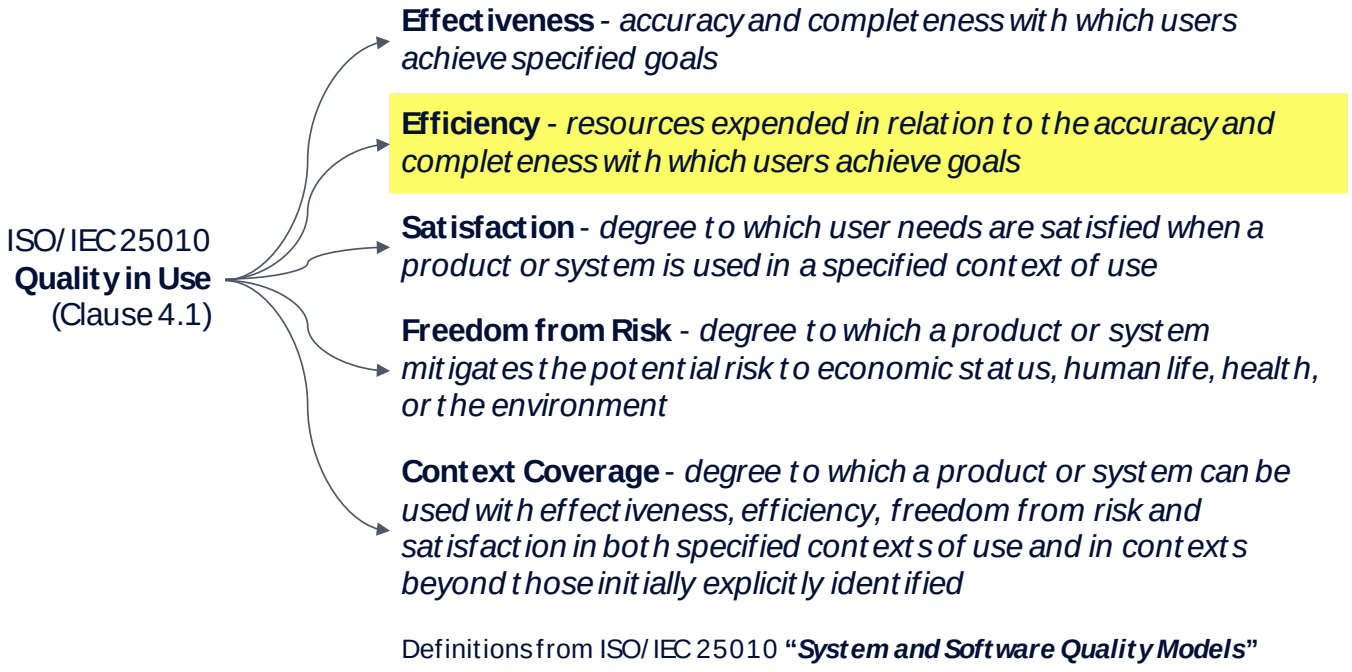

Figure 3-2 Efficiency as a Quality-in-Use Attribute

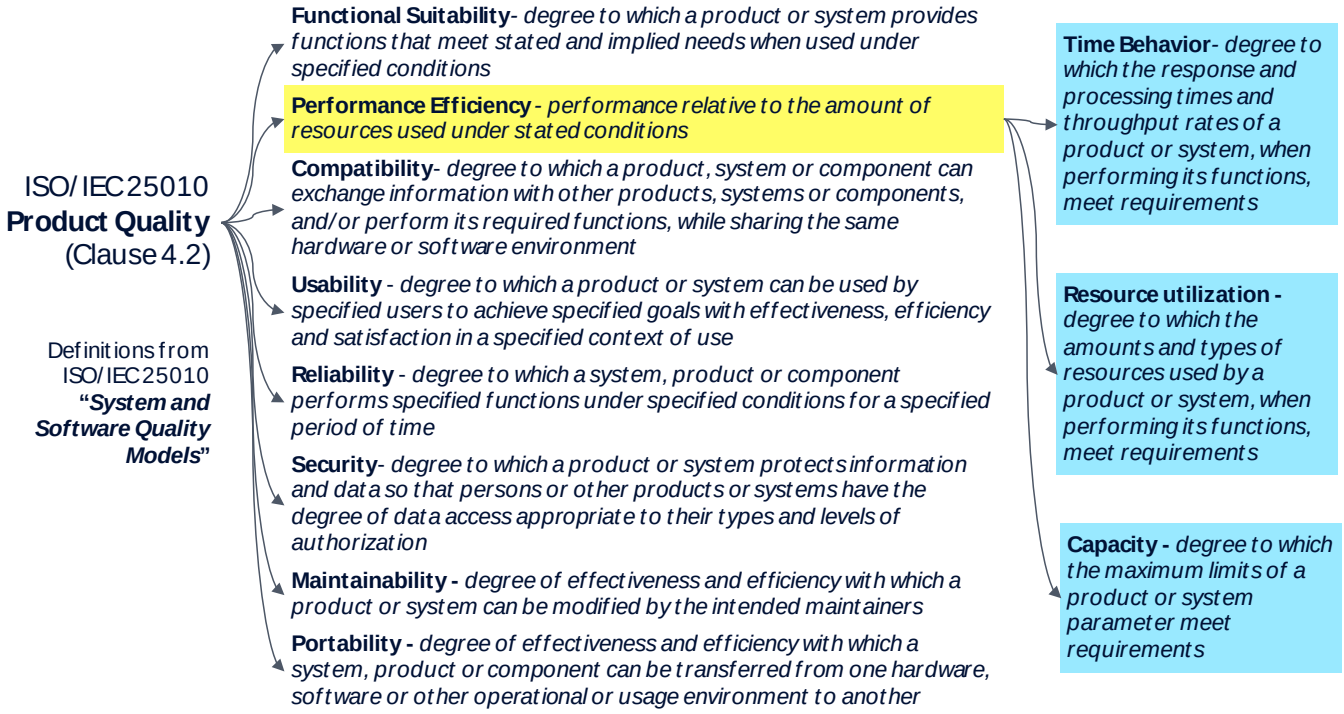

Figure 3-3 Performance Efficiency as a Quality Attribute

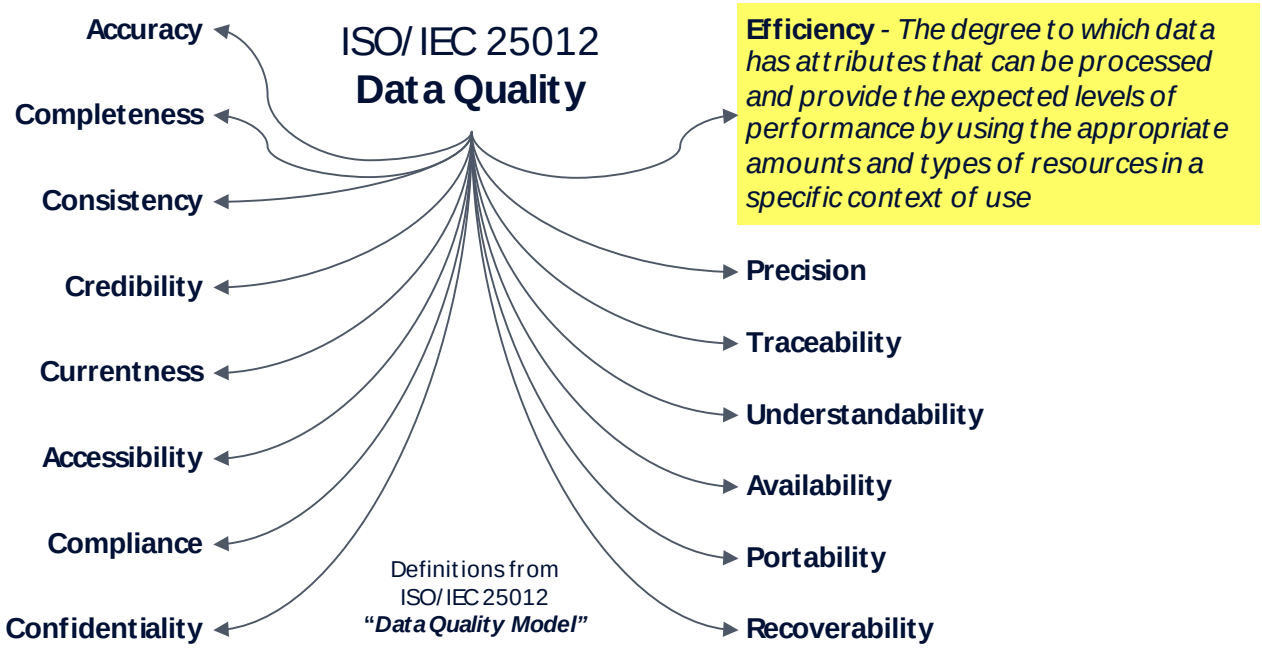

Figure 3-4 Efficiency as a Data Quality

Unfortunately, none of the ISO/IEC standard quality in use, product quality and data quality efficiency characteristics adequately addresses the *Target Use Cases* of section 2. NFV efficiency measurements to serve the *Target Use Cases* of section 2 should be based on defined subcharacteristics and characteristics that fit into one or more existing or new standardized quality model(s).

One should also recognize that several efficiency-related measurements have already been defined for NFV:

- ✓ **NFV Service Quality Metrics** – ETSI NFV INF-010 (ETSI, 2016-01) offers the following efficiency-related measurements:
  - VM and VN Provisioning Latency – relates to time efficiency
  - VM and VN Provisioning Reliability, VM Dead-On-Arrival (DOA) – provisioning-related failures wastes effort to detect, localize and resolve problems which negatively impacts efficiency

- ✓ **Automated Lifecycle Management Measurements** – QuEST Forum's Quality Measurement of Automated Lifecycle Management Actions offers the following efficiency-related measurements:
  - On Time Service Delivery – relates to time efficiency of automated lifecycle management action
  - Service Quality – automated actions that are not right-first-time waste effort to detect, localize and resolve problems which negatively impacts efficiency.

4 Why Is Efficiency More Complicated for NFV?

Efficiency is a more complex problem for NFV service components than for traditional elements for the following reasons:

- **Dynamic and variable levels of production** – the key cloud characteristic of rapid elasticity and scalability assures that "*physical or virtual resources can be rapidly and elastically adjusted, in some cases automatically, to quickly increase or decrease resources*"(ISO/IEC, 2014-10-15), so assuming fixed capacity configurations, resource consumption and service demand is unlikely to reflect actual operation.

- **Stochastic (aleatoric) uncertainty in virtualized service quality and throughput** – the key cloud characteristics of pooled resources and multi-tenancy, as well as the software technology that enables cloud service providers to offer virtualized infrastructure to cloud service customers introduces stochastic uncertainties in the consistency of delivered infrastructure service quality and throughput. Variations in resource service quality and throughput (especially episodes of lower throughput) will prompt cloud service customers to carry more online spare capacity to mitigate episodes of low infrastructure throughput . The more frequent the episodes of low infrastructure throughput are the more likely the cloud service customer is to have to carry excess spare resource capacity (thereby lowering effective efficiency) to mitigate the risk to service quality.

- **Explicitly decoupling software from hardware via virtualization increases the likelihood of imperfect mapping of application demands to the physical hardware capabilities**. Just as an operating system consumes some resource overhead to share time slices of a computer's physical compute, memory, storage and networking with individual processes, NFV infrastructure software consumes some resource overhead to share slices of virtualized compute, memory, storage and networking with VNF components across the Vn-Nf service boundary. Fairly and properly attributing the inevitable inefficiency due to imperfect mapping between application demand and physical hardware capabilities is important to drive efficiency improvement.

5 Modeling VNF Resource Usage

Figure 5-1 shows the high level NFV architecture from ETSI GS NFV-INF 001 *Infrastructure Overview* (ETSI, 2015-01). Notice that there are two external reference points between VNFs and the underlying NFV infrastructure:

- **[Vn-Nf]/VM** – highlighted via green outline on Figure 5-1 – described in (ETSI, 2015-01) as "*This reference point is the virtual machine (VM) container interface which is the execution environment of a single VNFC instance*."

- **[Vn-Nf]/N** - highlighted via blue outline on Figure 5-1 – described in (ETSI, 2015-01) as "*This reference point is the virtual network (VN) container interface (e.g. an E-Line or E-LAN) which carrying communication between VNFC instances. Note that a single VN can support communication between more than a single pairing of VNFC instances (e.g. an E-LAN VN).*"

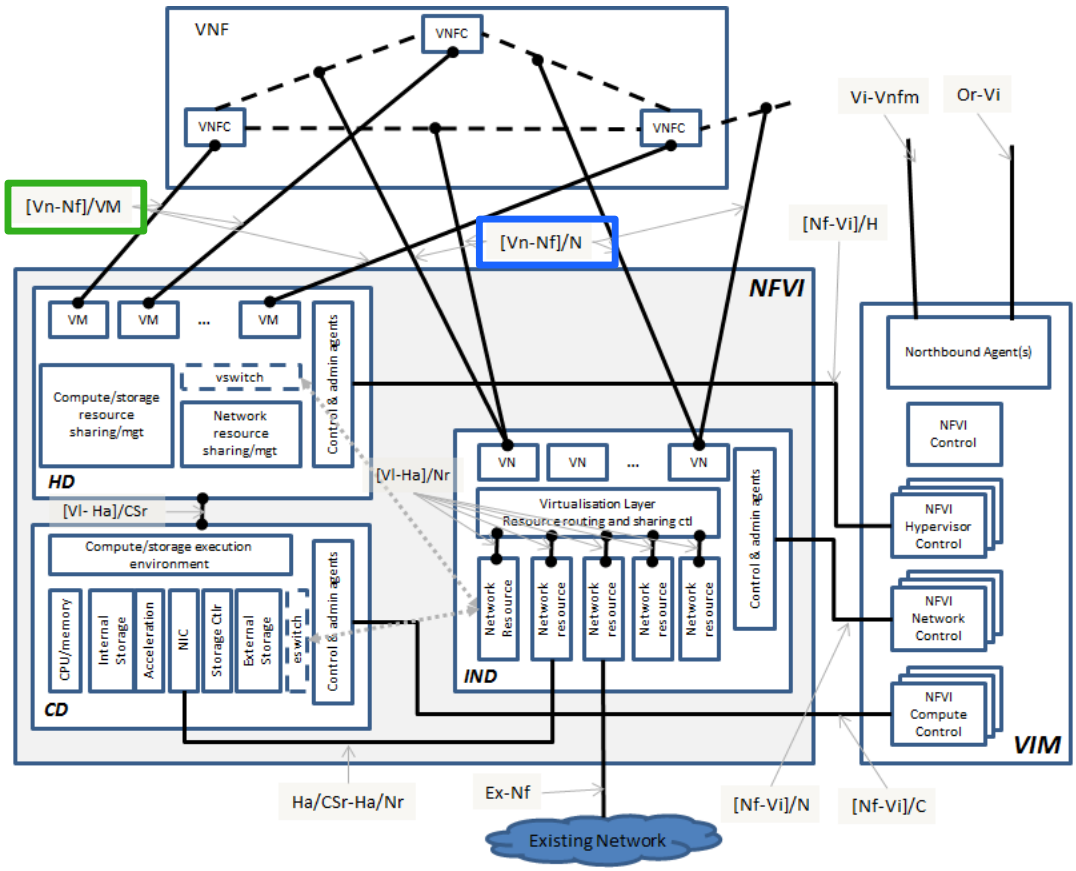

Figure 5-1 High Level Overview of NFV Architecture (from Figure 23 (ETSI, 2015-01))

Figure 5-1 illustrates that [Vn-Nf]/VM resources are fundamentally quantized as discrete resources with a 1:1 mapping between VNFC and VM. Note that [Vn-Nf]/VM infrastructure resources are allocated in whole units (e.g., 1 virtual machine with 4 virtual CPUs) rather than more fluid resources where an arbitrary amount of material can be obtained (e.g., 12.345 gallons of gasoline or 2.45 pounds of bananas). Most VNFs engineered with a modest, fixed application overhead for operations, management visibility and management controllability and an elastic pool of service components which deliver useful output. Each unit of useful application output is served by resource increments (e.g., virtual CPU instances) $\boldsymbol{Resource}_{Grow}$ with offer an increment of useful application output $\boldsymbol{Capacity}_{Grow}$; application overhead is served by resource overhead $\boldsymbol{Resource}_{Overhead}$. Note that each application has some $\boldsymbol{Capacity}_{Minimum}$ which is the largest capacity that can be served by the smallest supported production configuration of the VNF and some $\boldsymbol{Capacity}_{Maximum}$ which is the capacity that can be served by the largest supported production configuration.

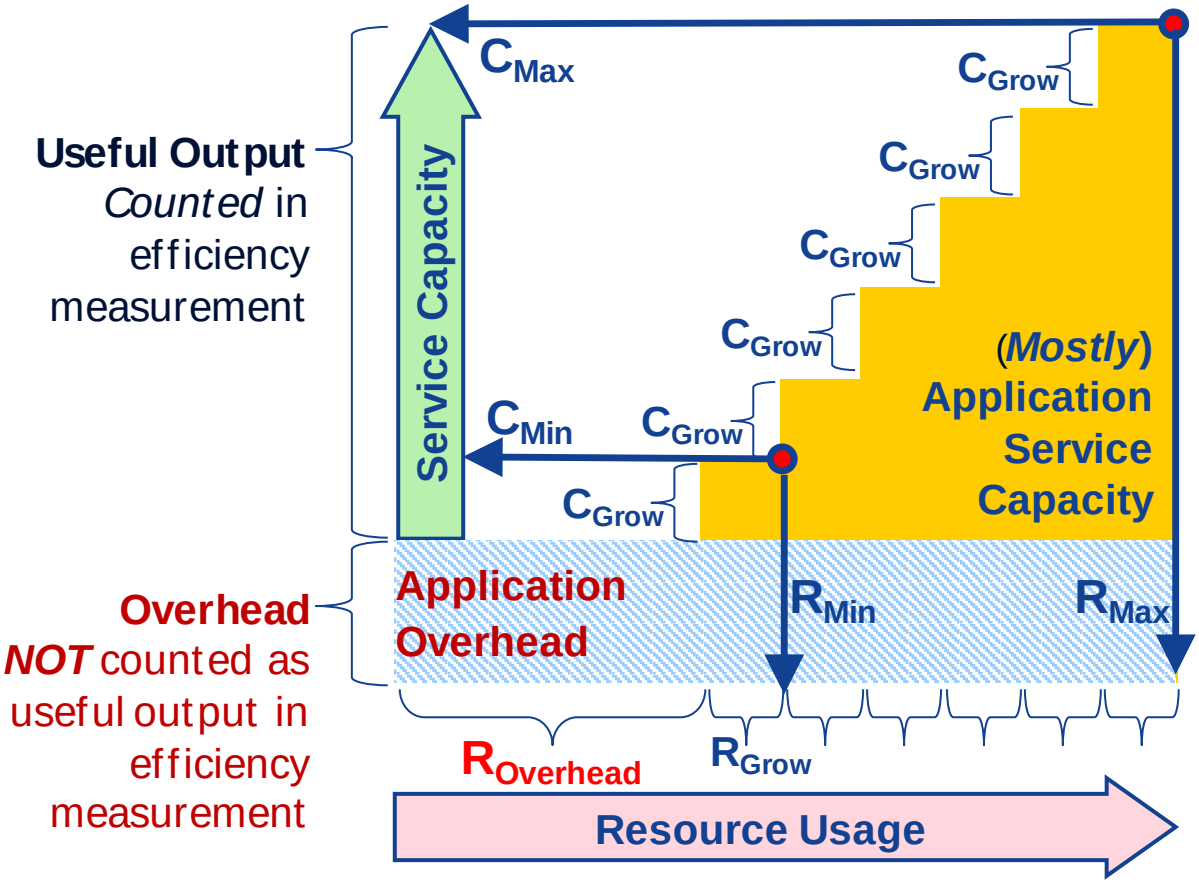

Figure 5-2 Canonical VNF Resource Commitment versus Service Capacity

6 Quantitatively Measuring Efficiency Qualities

To assure that end-to-end NFV efficiency measurements use consistent terminology and are rigorously specified, those measurements should conform to ISO/IEC 15939 "*Systems and Software Engineering — Measurement Process*," which is the most authoritative reference on quantitatively measuring qualities of software-based systems.

6.1 Measurement Information Model

Figure 6-1 overlays standard descriptions onto "*Figure A.1 — Key relationships in the measurement information model*" from ISO/IEC 15939**Invalid source specified.**.

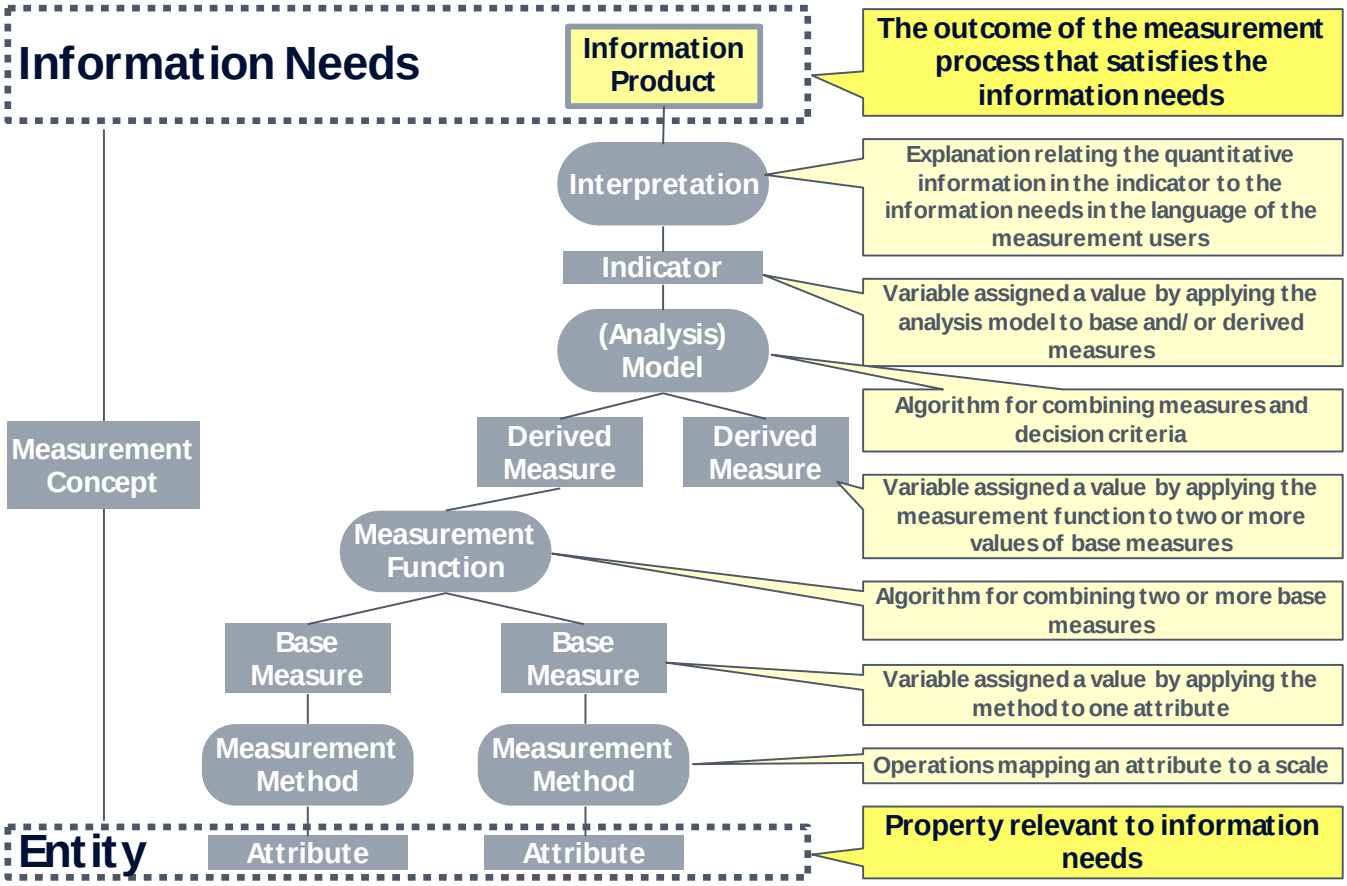

Figure 6-1 ISO/IEC 15939 Measurement Information Model (from Figure A.1 Invalid source specified.**)**

Figure 6-2 generically maps NFV efficiency measurements into the ISO/IEC 15939 measurement information model:

- Objective and quantitative measurement of an NFV service component's efficiency is the ***information product*** which serves the ***information needs*** of the *Target Use Cases* of section 2
- Target NFV service component is the ***entity*** being measured.
- NFV efficiency is the ***measurement concept***
- The measurement ***(analysis) model*** mathematically relates the IT service output produced by the target component to the resource input consumed by the target component
- The measurement combines ***derived measures*** of:
    1) Service output produced by the target service component
    2) Resource input consumed by the target service component
- Each of the *derived measures* is the result of some ***measurement function*** which applied a ***measurement method*** to some attribute of the target service component to produce some ***base measure***.

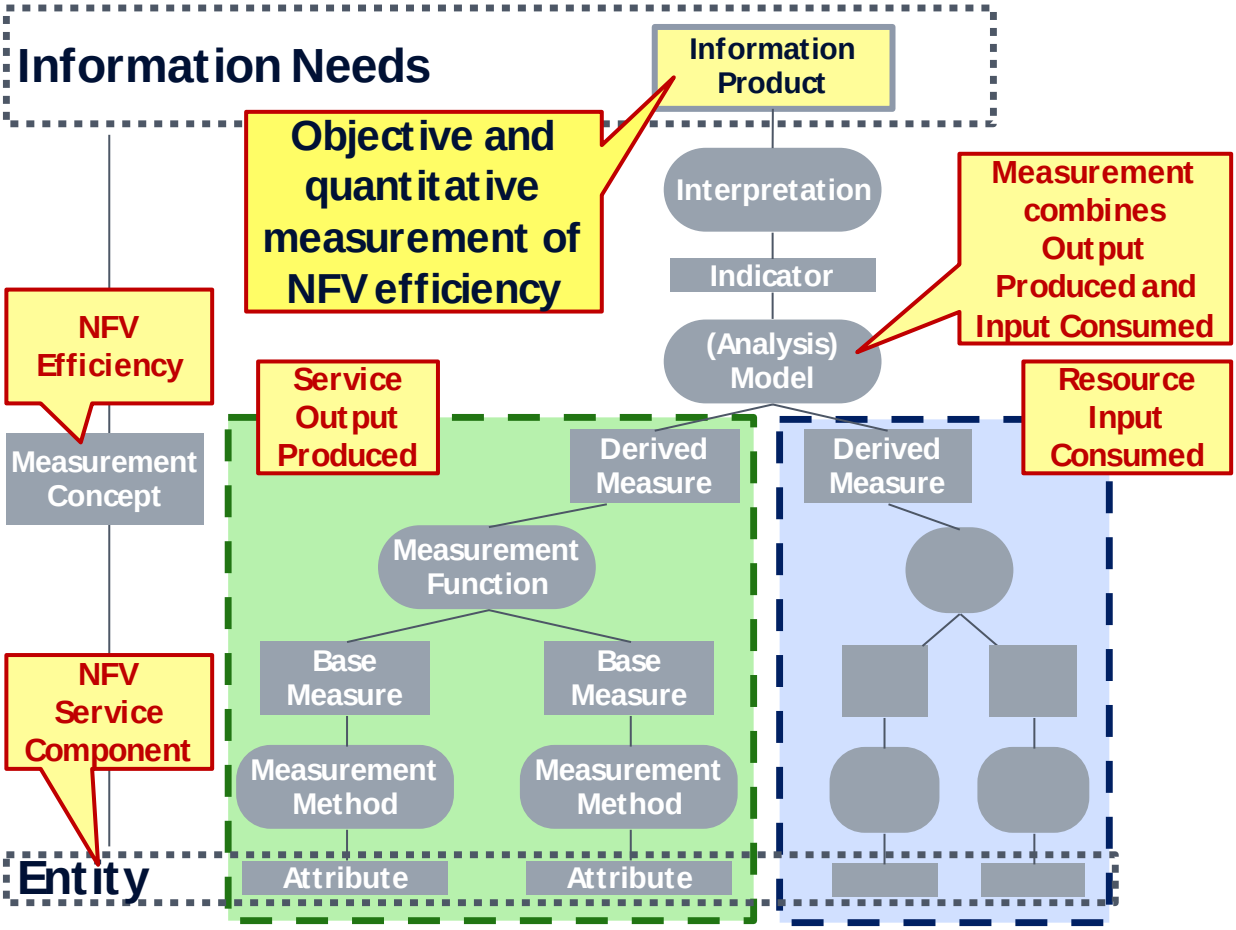

Figure 6-2 NFV Efficiency Measurement in ISO/IEC 15939 Measurement Information Model

NFV efficiency measurements should be specified as ISO/IEC 15939 measurement information models.

6.2 Standard Criteria for Selecting Measurements

Annex C of ISO/IEC 15939 offers many sample criteria for selecting measurements. Criteria from Annex C most applicable when selecting NFV efficiency measurements are:

1. ***sensitivity to context*** *…*
2. ***ease of interpretation*** *by measurement users and measurement analysts;*
3. ***relevance*** *to the prioritized information needs;*
4. ***ease of data collection****;*
5. *The* ***costs of collecting, managing, and analysing the data*** *at all levels should also be considered* (ISO/IEC, 2008-10-01)

7 VNF Efficiency

VNF efficiency is considered via the following sections:

- **VNF Resource Efficiency** (section 7.1) – characterizes the virtual resources consumed to serve a particular steady-state VNF workload; this leads to offers recommendation I - **Define VNF Resource Efficiency Measurement** (see section 12 *Recommendations)*.

- **VNF Elasticity Efficiency** (section 7.2) – quantitatively expresses how close a VNF's configured application capacity tracks to the 'perfect' capacity desired by the cloud service customer; this leads to offers recommendation **II - <u>Define VNF Elasticity Efficiency Measurement</u>** (see section 12 *Recommendations)*.

- **VNF Lifecycle Management Automation Efficiency** (section 7.3)

7.1 VNF Resource Efficiency

Service output of a VNF is inherently application dependent; for example, the application workload of a TL 9000 product category 1.2.8 ***session & network controller*** is measured differently from a TL 9000 product category 2.3 ***home location register*** or a TL 9000 product category 2.5 ***session border controller***. Thus, TL 9000 product-category-specific rules should be developed for quantifying the workload output of each NFV service component that will be covered by NFV efficiency measurements.

The following consumable <u>*virtual*</u> resource inputs to VNFs can be considered:

- Virtual CPU cores
- Virtual RAM
- Persistent storage
- Network throughput
- I/O throughput

As shown in Figure 7-1, VNF resource efficiency is characterized by two points:

- ✓ $\mathbf{VNF_{Max}}$ for the VNF service capacity of the largest supported configuration divided by the resources allocated for that configuration

- ✓ $\mathbf{VNF_{Min}}$ for the VNF service capacity of the smallest supported configuration divided by the resources allocated for that configuration

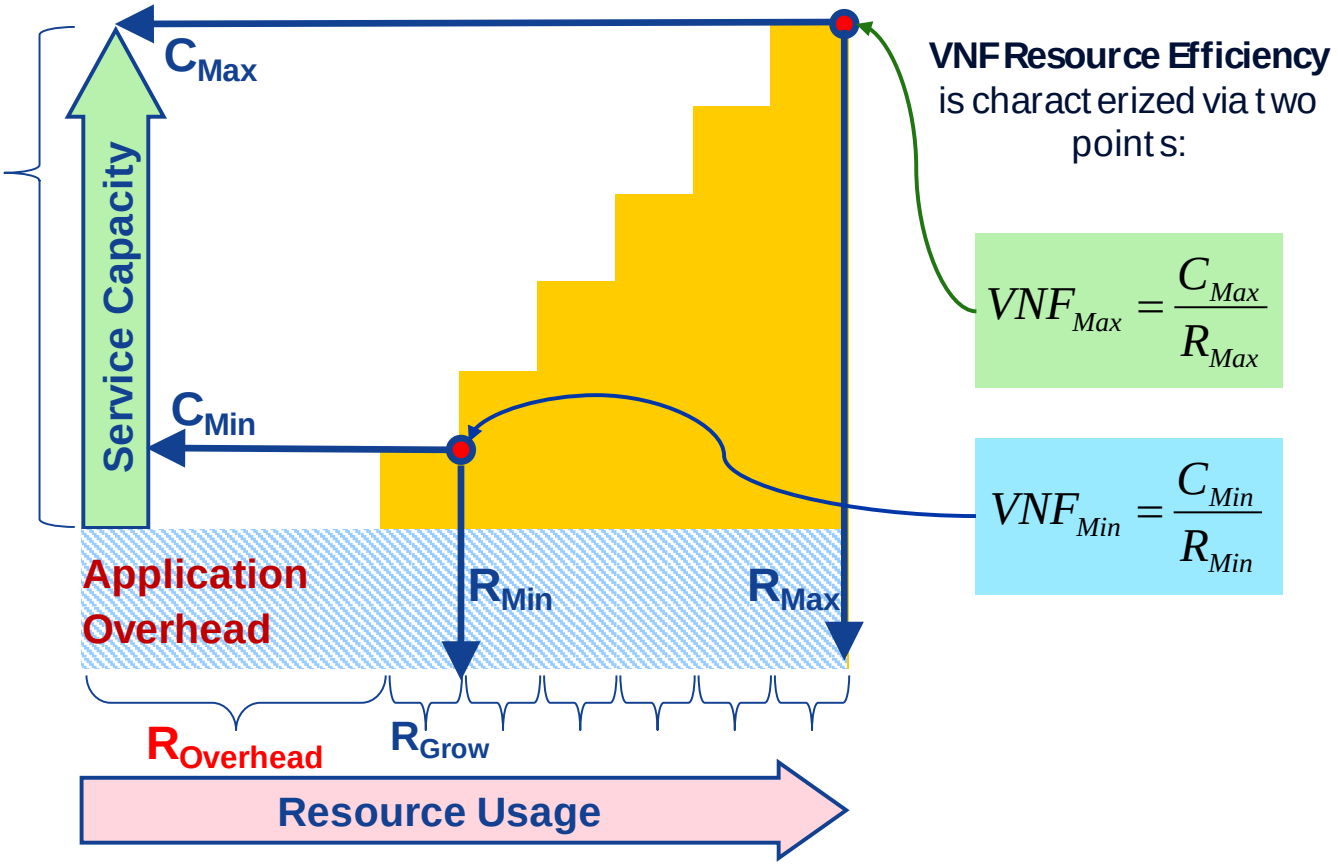

Figure 7-1 VNF Resource Efficiency Measurements

Figure 7-2 shows how resource efficiencies for VNFs from three different suppliers (X, Y and Z) could be visualized by plotting both $\boldsymbol{VNF_{Min}}$ and $\boldsymbol{VNF_{Max}}$ points for each supplier's VNF on a single chart. Note that the ***Min*** and ***Max*** points are deliberately connected by a dashed line because *VNF Elasticity Efficiency* (section 7.2) determines how smooth (versus jagged) the VNF resource efficiency function actually is.

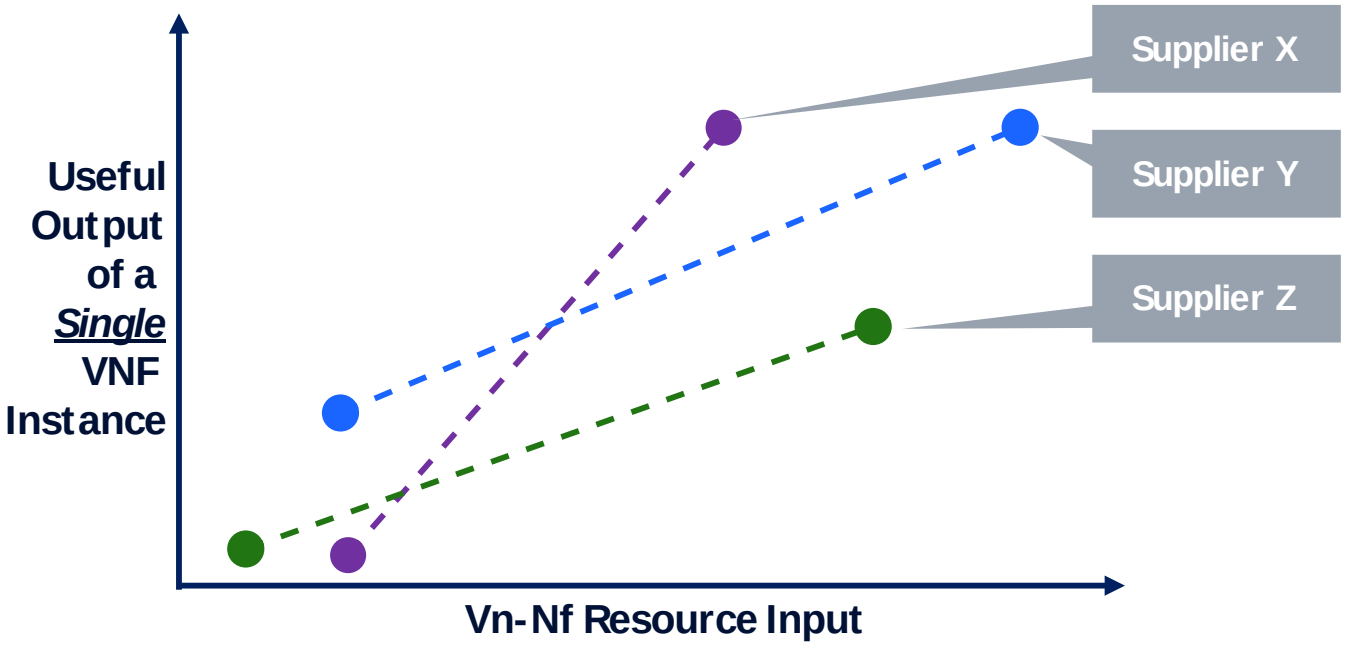

Figure 7-2 Comparing VNF Resource Efficiencies

Section 12 offers recommendation I - **Define VNF Resource Efficiency Measurement**.

7.2 VNF Elasticity Efficiency

One of six key characteristics of cloud computing is:

> ***Rapid elasticity and scalability****: A feature where physical or virtual resources can be rapidly and elastically adjusted, in some cases automatically, to quickly increase or decrease resources…*(ISO/IEC, 2014-10-15).

Elasticity of smoothly configurable resource services (e.g., electric power generation) can be measured as simple ramp rates (e.g., megawatts per minute). In contrast, Figure 7-3 visualizes VNF capacity changes due to discrete configuration change actions, such as adding or removing a VNFC instance to/from a pool of fungible worker component instances. Thus, the discrete, dynamic operational elasticity measurement considers the change in application capacity (as the delta-Y) and the lead time to fulfill the change (as the delta-X). As both (delta-X) lead time and (delta-Y) application capacity change get smaller and smaller it becomes easier for cloud service customers to instantaneously and perfectly match online resources allocated (which is likely to be correlated with pay-as-you-go costs) to instantaneous application demand (which is likely to be correlated with revenue).

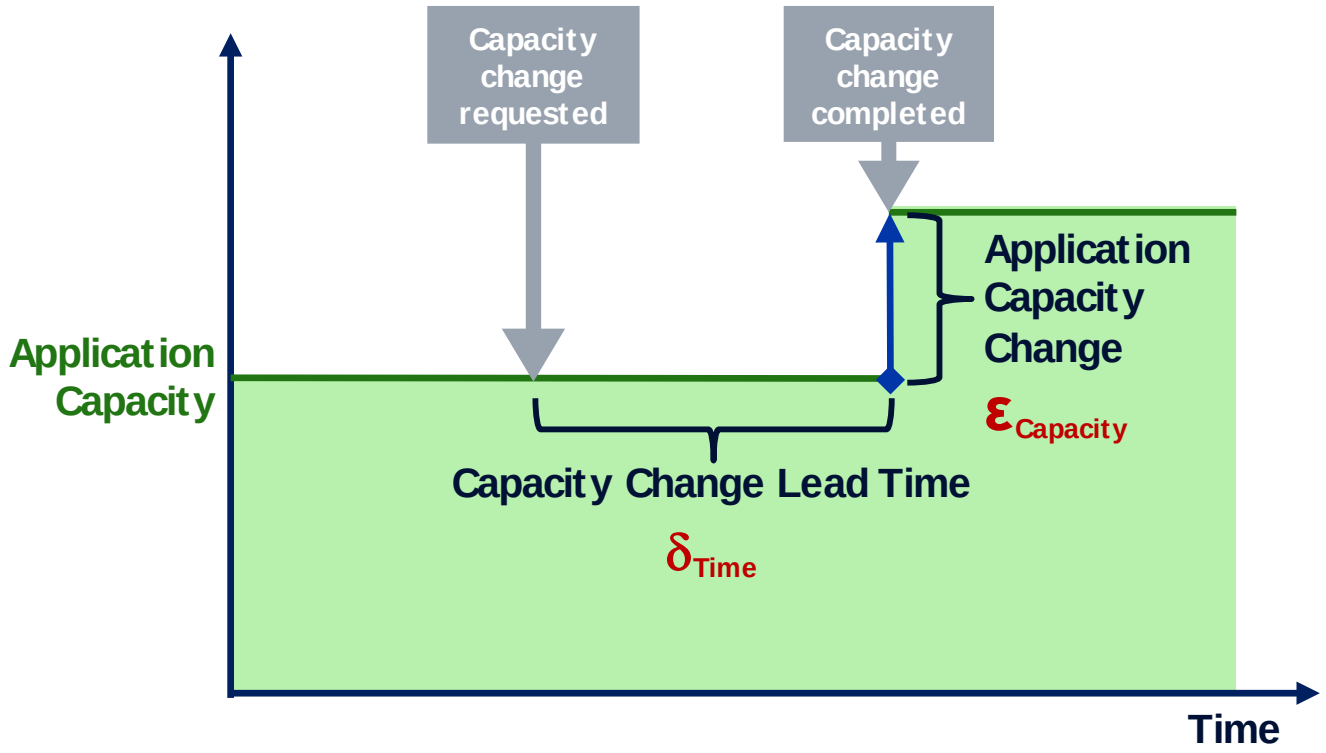

Figure 7-3 Dynamic Operational Efficiency of Cloud-Based Applications

Note that ***ElasticityEfficiency*** is ***expressed as pair of discrete numbers*** (*ApplicationCapacityChange* as service output and *CapacityChangeLeadTime* as consumed resource input) rather than a fraction because cloud elasticity actions are fundamentally non-instantaneous, discrete events.

$$ElasticityEfficiency = ApplicationCapacityChange : CapacityChangeLeadTime$$

As shown in Figure 7-4, ***ApplicationCapacityChange*** drives how jagged the actual VNF resource efficiency function is by characterizing how large the resource input : useful capacity output steps are between the $\boldsymbol{VNF_{Min}}$ and $\boldsymbol{VNF_{Max}}$ points on a resource efficiency map. ***CapacityChangeLeadTime*** is the average time to complete capacity growth actions from $\boldsymbol{VNF_{Min}}$ to $\boldsymbol{VNF_{Max}}$.

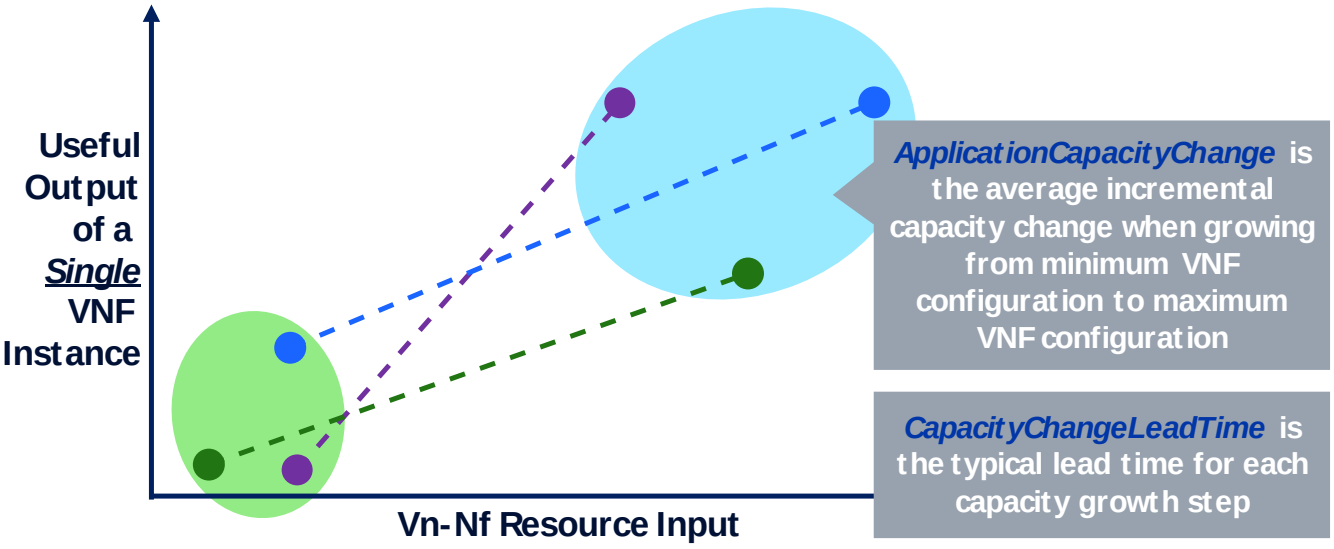

Figure 7-4 Elasticity Attributes on a VNF Resource Efficiency Map

Note that actual capacity change lead time includes latency directly attributed to VNF architecture as well as latency attributed to the cloud service provider for management, orchestration and infrastructure operations, and perhaps some cloud service customer policy-related interval to synchronize with enterprise and service management systems, complete automated validation testing and apply user traffic to newly grown service capacity. A formal efficiency measurement definition must precisely clarify these details.

Also note that VNF elasticity efficiency measures the characteristic as-built by the VNF supplier rather than the behavior as-configured by the cloud service customer. For example, a VNF supplier may engineer their VNF to grow service capacity via small (e.g., 1 vCPU) VNFCs, but for business reasons a cloud service customer may opt to configure their VNF deployment to grow service capacity in fewer, larger steps via medium (e.g., 4 vCPU) VNFCs. Obviously, the elasticity efficiency is worse when the VNF is configured to grow by 4 vCPU medium VNFCs rather than by 1 vCPU small VNFCs, but fewer, larger capacity change steps may yield better business results for a particular cloud service customer.

Section 12 offers recommendation **II - Define VNF Elasticity Efficiency Measurement**.

7.3 VNF Lifecycle Management Automation Efficiency

NFV management and orchestration enables VNF lifecycle management actions to be automated, thereby enabling cloud service customers to reduce human effort for operations, administration, maintenance and provisioning (OAM&P) of VNFs compared to PNFs. Objective and quantitative measurement of a VNF's 'automation efficiency' should characterize the likely operational efficiency improvement relative to reduction in human effort for OAM&P. ETSI NFV MAN-001 *Management and Orchestration* (ETSI, 2014-12) Clause 7.2 covers "Interfaces Concerning Virtualized Network Functions":

- VNF Package Management – Clause 7.2.1
- VNF Software Image Management – Clause 7.2.2
- VNF Lifecycle Operations Granting – Clause 7.2.3
- VNF Lifecycle Management – Clause 7.2.4
- VNF Lifecycle Change Notification – Clause 7.2.5
- VNF Configuration – Clause 7.2.6
- VNF Performance Management – Clause 7.2.7
- VNF Fault Management – Clause 7.2.8

Once the industry reaches consensus on the most important ratio of useful output to resource input for VNF lifecycle management automation, one or more VNF lifecycle management efficiency measurements can be developed.

8 NFV Infrastructure Efficiency

Figure 8-1 visualizes physical resource efficiency at the highest level as virtual compute, memory, storage and networking service output divided by physical resource inputs consumed by the target NFV infrastructure equipment. Service output of NFV infrastructure is delivered across the Vn-Nf service reference point.

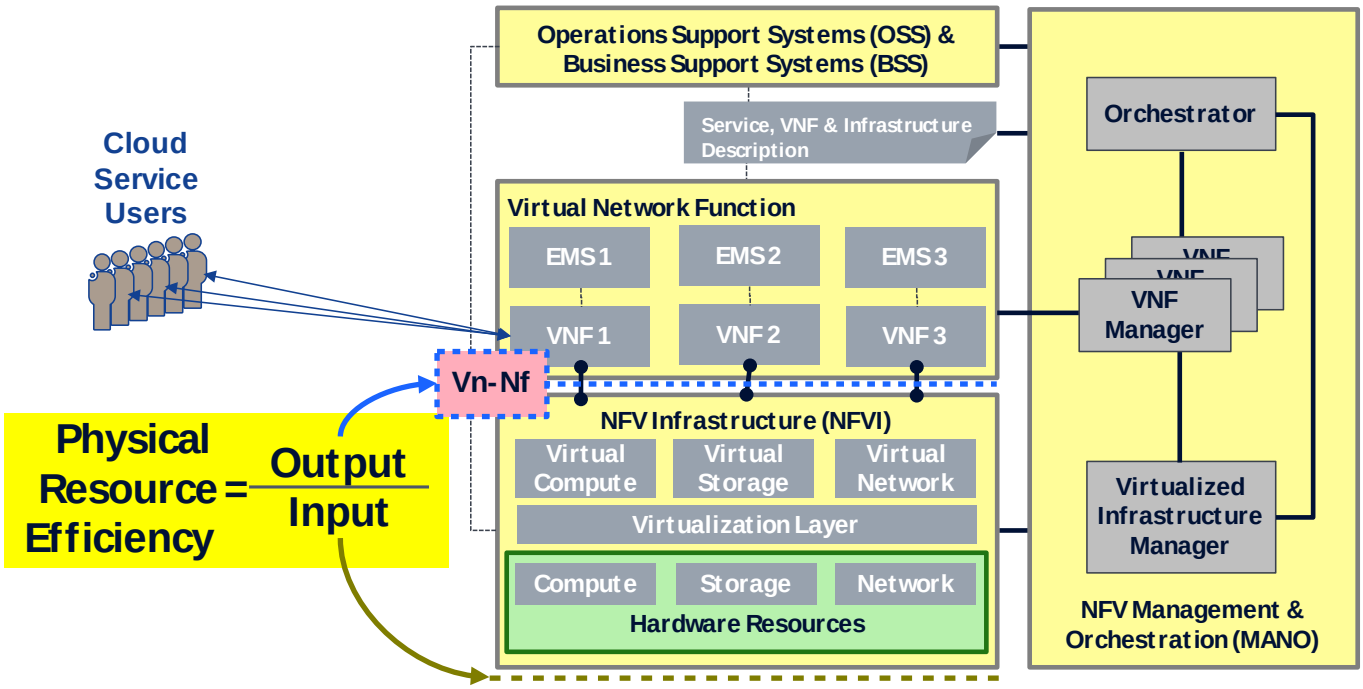

Figure 8-1 Physical Resource Efficiency of NFV Infrastructure

The topic is considered in the following sections:

- **Energy Efficiency of NFV Infrastructure** (section 8.1) derives recommendation **III - Leverage NFV-Related Energy Efficiency** (see section 12 *Recommendations)*.
- **Virtualization Efficiency of NFV Infrastructure** (section 8.2)

8.1 Energy Efficiency of NFV Infrastructure

ETSI ES 201 554 "*Measurement method for Energy efficiency of Core network equipment*" (ETSI, 2012-04) offers the following definitions:

> ***energy efficiency***: *relation between the useful output and energy consumption.*
>
> ***useful output****: maximum capacity of the system under test which is depending on the different functions*
> *NOTE: It is expressed as the number of Erlang (Erl), Packets/s (PPS), Subscribers (Sub), or Simultaneously Attached Users (SAU).*
>
> ***energy consumption****: amount of consumed energy*
> *NOTE: It is measured in Joule or kWh (where 1 kWh = 3,6 × 106 J) and corresponds to energy use.*

From Clause 4.3 *Power Consumption* of (ETSI, 2012-04):

> *The load levels are defined as:*
> - *Specification: T_S - the maximum capacity according to the vendor's specification of the specific implementation of the function*
>   - *High: $T_H = 1{,}0 \times T_S$*
>   - *Mid: $T_M = 0{,}7 \times T_S$*
>   - *Low : $T_L = 0{,}1 \times T_S$*
>
> …
> *The power consumption levels associated with the above load levels are defined as:*

- *High: P_H = average power consumption [W] measured at T_H*
- *Mid: P_M = average power consumption [W] measured at T_M*
- *Low: P_L = average power consumption [W] measured at T_L*

The average power consumption is defined as:

$$P_{avg} = \alpha \times P_L + \beta \times P_M + \gamma \times P_H \, [W]$$

Where α, β, and γ are weight coefficients selected such as $(\alpha + \beta + \gamma) = 1$.

Figure 8-2 from Clause 4.4 *Shaping of Weight Coefficients* illustrates how the standard coefficients for voice traffic relate to sample real world data.

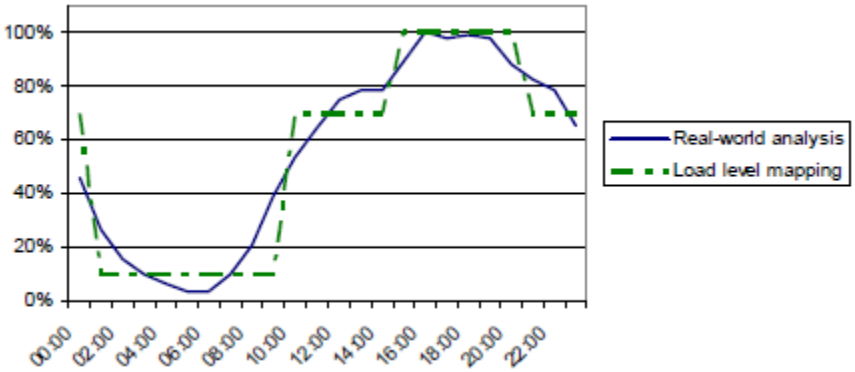

Figure 8-2 Load Level Mapping Overlaid on Real World Data (Figure 3 from (ETSI, 2012-04))

Clause 4.5 *Energy Efficiency* stipulates:

The Energy Efficiency Ratio metric, the comparable performance indicator, for Core networks is defined as:

EER = Useful Output / P_{avg} [Erlang/W | PPS/W | Subscribers/W | SAU/W]

Where Useful Output is the maximum capacity of the system under test (T_S) which, depending on the different functions, is expressed as the number of Erlang (Erl), Packets/s (PPS), Subscribers (Sub), or Simultaneously Attached Users (SAU). By using the defined traffic models, Useful Output can be translated to Subscribers (Sub) or Simultaneously Attached Users (SAU) also for functions which normally have the maximum capacity expressed in Erlang (Erl) or Packets/s (PPS).

Notice that the units of "Useful Output" (i.e., Erlangs, packets per second, subscribers or simultaneously attached users) for Energy Efficiency are explicitly tied to the nature of the target system.

Clause 4.1 "Black box" of ETSI ES 201 554 "*Measurement method for Energy efficiency of Core network equipment*" (ETSI, 2012-04) includes:

The system under test is seen as a "black box", i.e. only the total power consumed by the device or shelf/shelves is/are measured and not different parts of the device or shelf/shelves. A "black box" can be viewed solely in terms of its input,output and transfer characteristics without any knowledge of its internal workings.

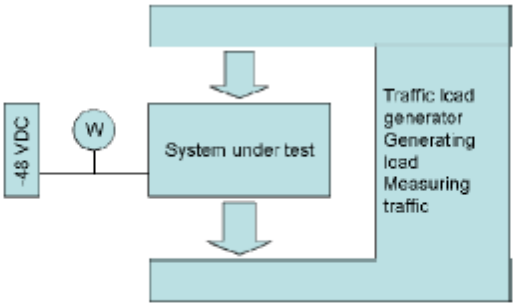

Figure 2: Measurement set-up of system under test

Figure 8-3 applies ETSI ES 201 554 Clause 4.1 energy efficiency measurement principles to NFV infrastructure.

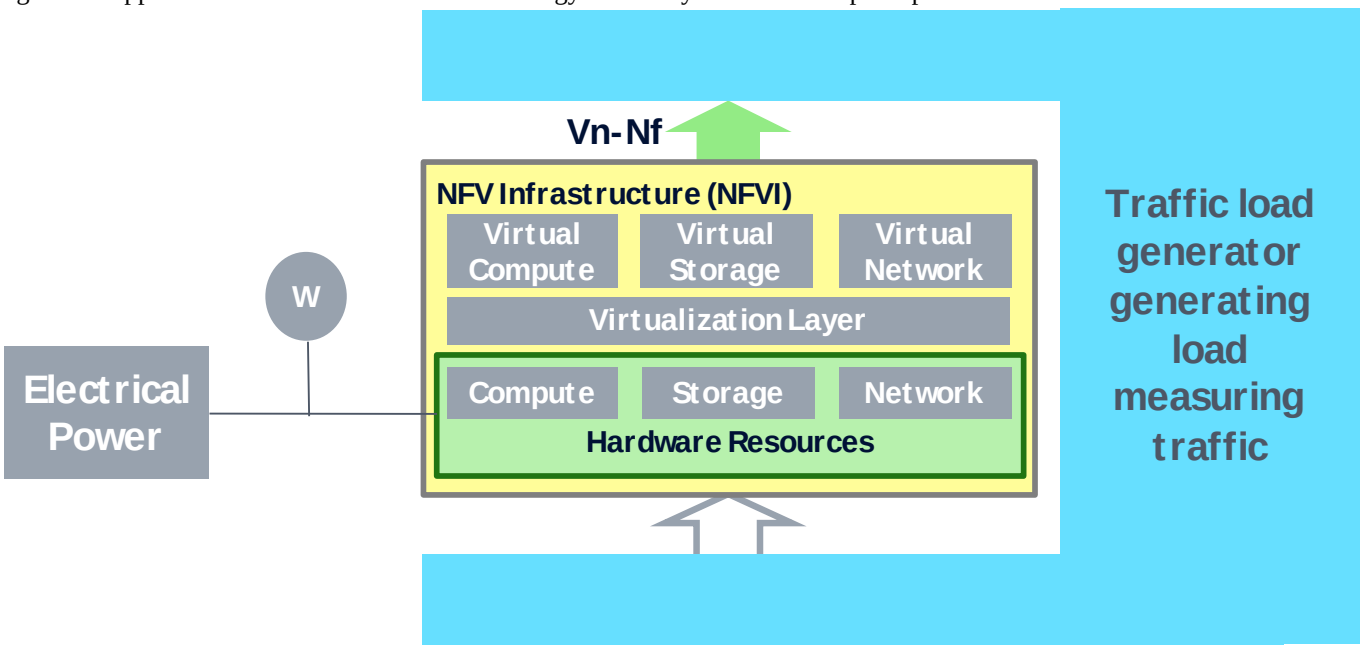

Figure 8-3 Example of Energy Efficiency Measurement Setup for NFV Infrastructure

Additional methods and principles for the assessment of mobile network energy efficiency are defined in ETSI ES 203 228**Invalid source specified.**.

Section 12 *Recommendations* includes ***III*** **- Leverage NFV-Related Energy Efficiency .**

8.2 Virtualization Efficiency of NFV Infrastructure

As shown in Figure 8-4, virtualization efficiency is the ratio of useful output of a benchmarking application running directly on bare metal hardware resources compared to the useful output when the same benchmarking application runs on NFV infrastructure hosted on the same hardware resources.

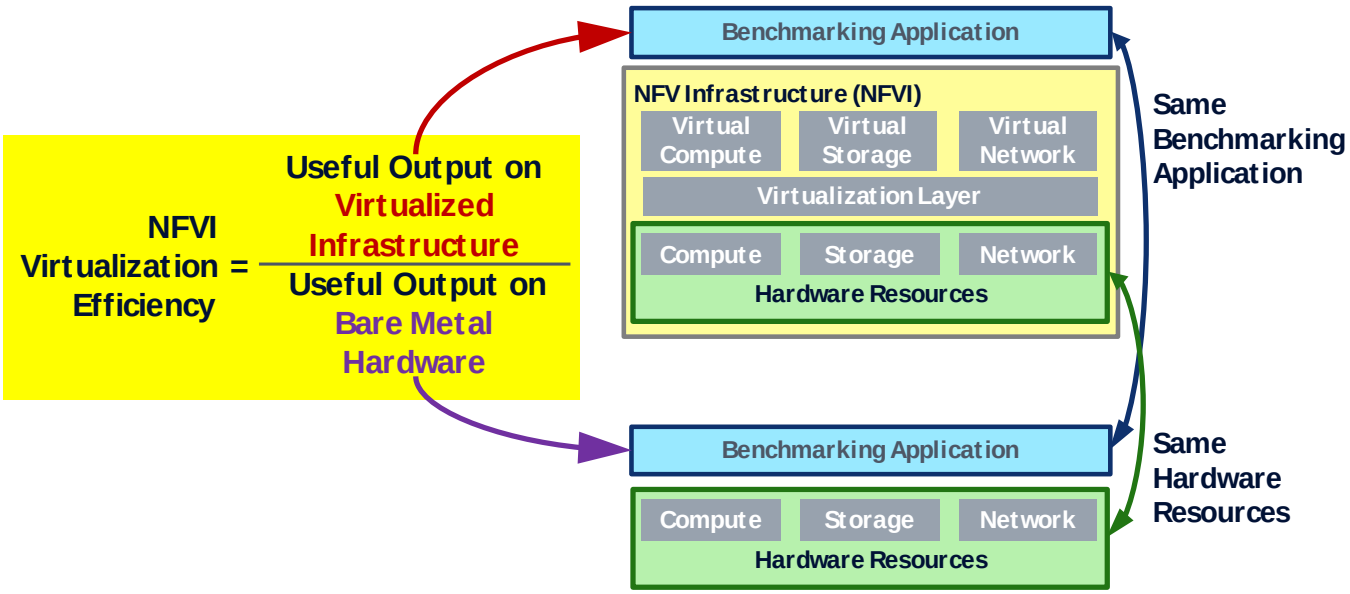

Figure 8-4 Virtualization Efficiency Measurement for NFV Infrastructure

While this is conceptually a simple measurement, results are likely to be highly sensitive to context, especially:

1. Target benchmarking application and application configuration
2. Specific hardware resources
3. Configuration of virtualization layer
4. Placement of application components into virtualized resources

Thus, this measurement may be useful as an *Workload Efficiency Measurements in Supplier* Selection (section 2.2) and perhaps for *Estimate Likely Operating Expenses of an NFV-based Service or Solution* (section 2.1), but it is not likely to be useful for *Workload Efficiency Measurements in Ongoing* Operations (section 2.3), *Workload Efficiency Measurements in Performance* SLAs (section 2.4) or *Benchmark Workload Efficiency* Performance (section 2.5).

9 NFV Management and Orchestration Efficiency

NFV management and orchestration offers valuable service output to:

- ✓ **Cloud service provider** – the useful service output for cloud service providers delivered by NFV management and orchestration is automating management of NFV infrastructure. Policy objectives of CSPs will undoubtedly vary between organizations, and perhaps across space and time. For instance, some CSPs may wish to minimize their carbon footprint while others may wish to offer best-in-class execution times for automated lifecycle management actions, while others seek to aggressively manage the yield/return on their infrastructure investment. Thus, it is unlikely that any single management and orchestration efficiency measurement would be important to all cloud service providers.

- ✓ **Cloud service customers** – the useful service output for cloud service customers provided by NFV management and orchestration is automating lifecycle management actions on behalf of CSCs. Policy objectives of CSC's will undoubtedly vary between organizations, and perhaps across space and time. For instance, some CSCs may wish to optimize geographic placement of resources to shorten network transport times, while other CSCs may wish to optimize placement to minimize their resource costs, and still others may have sophisticated policies related to failure group sizes, regulatory considerations or other factors, and yet others may simply want the fastest possible execution of automated lifecycle management actions. Thus, it is unlikely that any single NFV management and orchestration efficiency measurement would be important to all cloud service customers.

Once the industry reaches consensus on the most important ratio of useful output to resource input for NFV management and orchestration, one or more NFV efficiency measurements for management and orchestration can be developed.

Note that VNF Managers (VNFM) can be the responsibility of either the cloud service provider or the cloud service customer, so efficiency measurements should address both of these deployment options.

10 Evolving Energy Efficiency Measurement

NFV fundamentally decouples the hardware which actually consumes electrical power from the VNF software which utilizes the compute, memory, storage and networking services delivered by hardware across the Vn-Nf reference point. Thus, an energy efficiency measurement which relates useful service output like data volume, Erlangs or packets per second to an energy input like Joules or Watt hours must appropriately combine factors which have been explicitly decomposed by the NFV architecture. Figure 10-1 illustrates how an ETSI ES 201 554 energy efficiency measurement could be applied to NFV. The energy efficiency for a particular service (e.g., IMS) would be measured as useful service output (e.g. data volume in kbits) per Joule. Other examples for mobile network energy efficiency can be derived from ETSI ES 203 228.

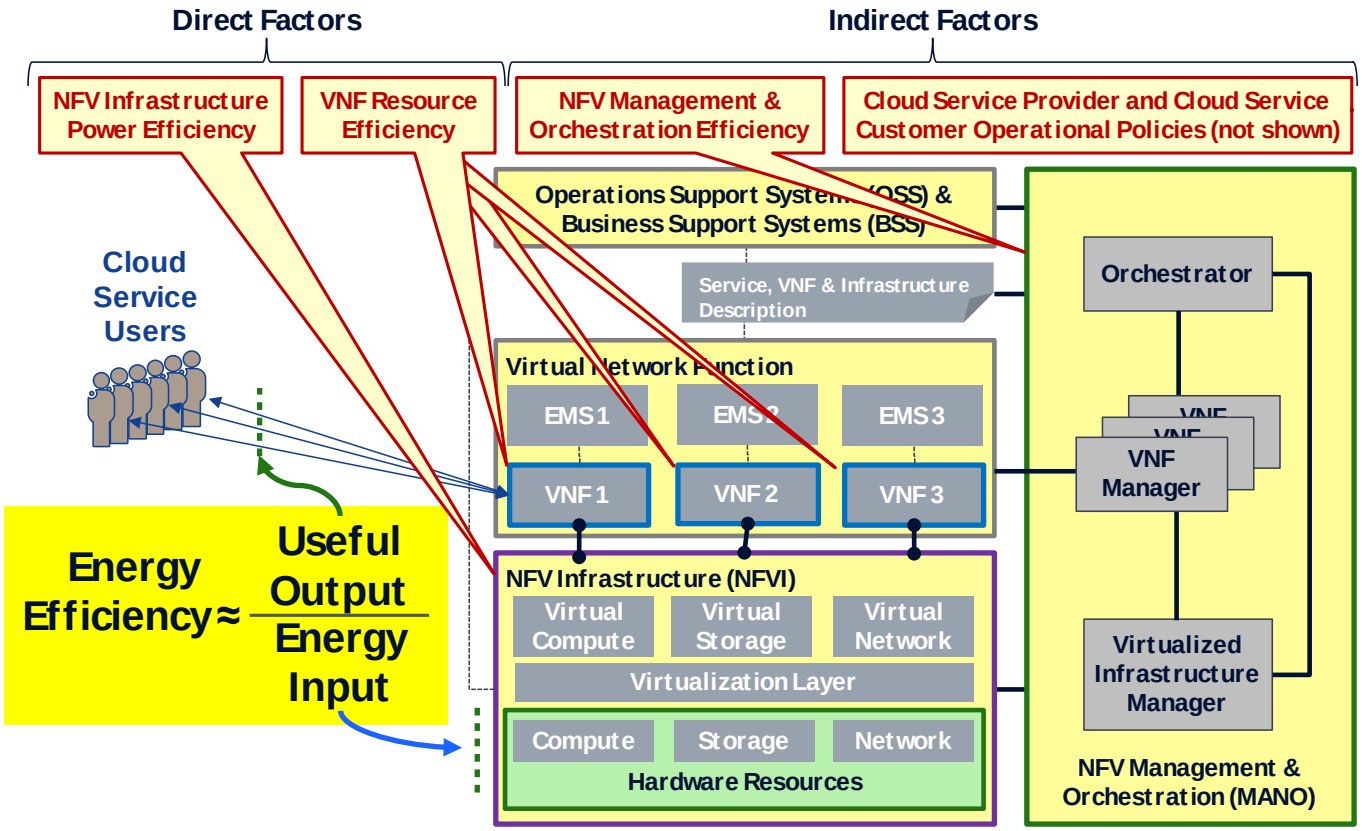

Figure 10-1 Example of Energy Efficiency in the Context of NFV

As shown in Figure 10-1, two indirect factors have a material impact on the actual NFV energy efficiency ratio observed by a service provider:

1. **NFV management and orchestration efficiency** – management and orchestration elements --- especially the orchestrator and virtualized infrastructure manager --- actually map resources serving VNF instances onto NFV infrastructure elements. Just as heap memory allocators and file systems can vary in how efficiently they allocate heap or disk resources, NFV management and orchestration elements can vary in how efficiently they allocate NFV infrastructure resources. Just as with heap memory allocators and file systems, sometimes one optimizes for speed (e.g., 'first fit') and sometimes one optimizes for minimum resource usage (e.g., 'best fit'). Thus, the appropriate 'useful output' to measure for NFV management and orchestration efficiency may vary between cloud service customers and cloud service providers based on their operational policies and business models.

2. **Cloud service provider and cloud service customer operational policies** – operational policies of the cloud service provider organization that owns and operates NFV infrastructure determines how aggressively power usage by infrastructure equipment will be managed, such as:
    a. How much spare infrastructure capacity is held online to rapidly serve resource allocation requests
    b. How aggressively VNF workloads are consolidated (e.g., via live migration and best fit resource allocations) to minimize infrastructure waste due to fragmentation
    c. How aggressively infrastructure equipment is powered on and off with varying application demand.

    Operational policies by the cloud service customers that operate VNFs impacts resource consumption, such as how much spare application is held online to mitigate failures and serve unforecast demand, and how much testing is done on new releases and newly grown application capacity before entering production.

Note that while energy efficiency is likely to be a concern for telcos, accountability for overall energy efficiency is fundamentally split between:

- **Cloud service customer organizations** which select and operate VNFs which consume Vn-Nf service of NFV infrastructure to deliver useful application service output

- **Cloud service provider organization** which select and operate NFV infrastructure, and supporting management and orchestration systems, which consume energy and deliver Vn-Nf service to cloud service customer organizations' VNFs

To efficiently drive continuous improvement, efficiency measurements should be appropriately aligned with the responsibilities of the organizations that own and operate both VNFs and NFV infrastructure.

11 Proposed Requirements for NFV Efficiency Measurements

Efficiency measurement definitions should conform to the following requirements:

1) **Workload efficiency metrics shall not include monetary resources** (i.e., *NOT* use currency as units of efficiency). The price that a customer pays for resource inputs shall be separate from the level of resources consumed by the target entity to produce the desired output. Not only are prices of input resources likely to vary across providers and geographies, but relative prices of resources may be different which may lead different customers to prioritize improvements in various efficiency measurements for the same VNF slightly differently.

2) **Workload efficiency metrics shall be specified per ISO/IEC 15939** "*Systems and Software Engineering - Measurement Process*"

3) **Counting rules for output of a target element shall vary by TL 9000 product category**. For instance, the counting rules of VNF output may be very different for a product category 2.3 home location register and a category 2.5 session border controller.

12 Recommendations

This paper recommends the following actions

- ✓ I - **Define VNF Resource Efficiency Measurement**
- ✓ II - **Define VNF Elasticity Efficiency Measurement**
- ✓ III - **Leverage NFV-Related Energy Efficiency**

I. Define VNF Resource Efficiency Measurement

An appropriate standards development organization should create a VNF Resource Efficiency Ratio measurement (see section 7.1 *VNF Resource Efficiency*) for the useful service output of a VNF as a function of NFV infrastructure resources (i.e., [Vn-Nf]/VM, [Vn-Nf]/N) consumed. Note that the specific measurement of VNF service output is driven by the primary functionality of the target VNF. For consistency with other Telecom quality measurements, VNF-specific measurement definitions of service output should be tied to the TL 9000 product category of the target VNF. In principle, the higher the VNF resource efficiency ratio, the lower the cloud service customer's usage of NFV infrastructure resource services.

II. Define VNF Elasticity Efficiency Measurement

An appropriate standards development organization should create a VNF Elasticity Efficiency Ratio measurement (see section 7.2 *VNF Elasticity Efficiency*) that objectively and quantitatively characterizes how effectively a VNF supports rapid elasticity and scalability. In principle, the higher the VNF automation efficiency ratio, the closer a VNF instance can track with the cloud service customer's notion of perfect application capacity.

III. Leverage NFV-Related Energy Efficiency Standardization

NFV Energy Efficiency assessment methods should be aligned with existing standards such as ETSI ES 201 554 "*Measurement Method for Energy Efficiency of Core Network Equipment*", ETSI ES 203 228 "*Assessment of mobile network energy efficiency*" as well as ETSI EN "*energy efficiency method and KPI for NFV applications*" (under development).

NFV Energy Efficiency assessment should be back-to-back with VNF resource efficiency so that Vn-Nf service output of produced by NFV infrastructure cancels out with Vn-Nf resource input to VNFs to estimate energy usage for user services offered by VNFs. In principle, the higher the NFV infrastructure energy efficiency, the lower the cloud service provider's usage of electric power.

13 Works Cited

Axelos Limited. (2011). *ITIL® glossary and abbreviations.* Axelos Limited.
ETSI. (2012-04). *ES 201 554 - Measurement Method for Energy Efficiency of Core Network Equipment.* Sophia Antipolis: European Telecommunications Standards Institute.
ETSI. (2015-01). *GS NFV-INF 001 Network Functions Virtualization Infrastructure Overview.* Sophia Antipolis: European Telecommunications Standardization Institute.
ETSI. (2014-12). *NFV MAN-001 Management and Orchestration.* Sophia Antipolis, France: European Telecommunications Standardization Institute.
ETSI. (2016-01). *NFV-REL 005 - Network Function Virtualisation Report on Quality Accountability Framework.* Sophia Antipoils Cedex, France: European Telecommunication Standardization Institute.
ISO/IEC. (2008-10-01). *15939 - Systems and Software Engineering - Measurement Process.* Geneva: International Organization for Standardization and International Electrotechnical Comission.
ISO/IEC. (2014-10-15). *17788 - Cloud computing -- Overview and vocabulary.* International Organization for Standardization & International Electrotechnical Committee.
ISO/IEC. (2005-08-01). *25000 Software product Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE.* Geneva: International Organization for Standardization.
ISO/IEC. (2011-03-01). *25010 System and Software Quality Models.* Geneva: International Organization for Standardization.
ISO/IEC. (2008-12-15). *25012 Data Quality Model.* Geneva: International Organization for Standardization.
QuEST Forum. (2012-12-31). *TL 9000 Measurements Handbook Release 5.0.* QuEST Forum.

14 Annex A - TM Forum Operational Efficiency Metrics

Table 14-1 enumerates the operational efficiency metrics defined by TM Forum in GB 988, although none of them directly address NFV workload efficiency.

Table 14-1 TM Forum GB 988 Operational Efficiency Metrics

name	egid
% Problem Report Closed, By Cause Type	A-OE-3a
% Repair Time, Of Repair And Maintenance Time	A-OE-3b
Mean Time Between Failures	A-OE-3c
% Billing Error Cost, Of Revenue Billed	B-OE-3a
% XDR Falling Into Suspense	B-OE-3c
% Collectable Debt Written Off, Of Revenue Collected	B-OE-3d
% Bill Value Unpaid	B-OE-3e

% PrePaid Customer Erroneously Identified As PostPaid	B-OE-3g
% Order Requiring Rework, By Cause Type	F-OE-3a
# Hours Fulfillment Issue Resolution Time, To Customer Acceptance, Per Fulfillment Issue	F-OE-3b
% Order Requiring Rework	F-OE-3c
% Order With Pending Error Fixes	F-OE-3d
# Minutes Problem Handling Time, To Service Restoration, Per Problem Report Closed	A-OE-2a
# Hours Service Problem Handling Time, Per Service Problem Resolved	A-OE-2b
Days Sales Outstanding	B-OE-2b
# Hours Order Fulfillment Time, From Activation, To Bill Dispatch, Per Order	B-OE-2c
# Hours Order Fulfillment Time, From Bill Dispatch, To Cash Received, Per Order	B-OE-2d
# Hours Customer Payment Handling Time, From Receipt, To Posted In Billing, Per Customer Payment	B-OE-2e
# Hours Billing Process Outage Time, Per Bill Processing Fault	B-OE-2f
# Hours Order Fulfillment Time, From Ordering, To Activation, Per Order Accepted By Customer	F-OE-2a
# Hours Order Fulfillment Time, From Ordering, To Activation, Per Order Accepted By Customer	F-OE-2b
# Hours Pricing Change Activity Time, Per Pricing Change	O-OE-2a
% Assurance Cost, Of Revenue	A-OE-1a
% Assurance Cost, Of Opex	A-OE-1b
% SLA Management Cost, Of Revenue	A-OE-1c
$ Assurance Cost, Per Service Problem Resolved	A-OE-1f
% Billing Cost, Of Revenue Billed	B-OE-1a
% Bill Requiring Manual Intervention	B-OE-1b
% Collections Cost, Of Revenue Billed	B-OE-1c
$ Billing Cost, Per Bill Produced	B-OE-1f
% Fulfillment Cost, Of Revenue For Services Newly Fulfilled	F-OE-1a
% Sales Cost, Of Revenue For Services Newly Fulfilled	F-OE-1b
% Fulfillment Cost, Of Opex	F-OE-1c
% Revenue, By Channel Type	F-OE-1d
% Revenue, By Channel Type, Of Channel Cost	F-OE-1e
$ Fulfillment Cost, Per Installation	F-OE-1f
# Problem Reports, Per NOC FTE Assigned To Problem Resolution	A-OE-6a
% Future Infrastructure Build Investment, Of Revenue	F-OE-6