Taxonomy of metrics for cloud services

APPENDIX 1. Classification of QoE metrics found in the systematic mapping

Legend	
Data Extraction criteria	Possible Answer of each data extraction criteria
QoS characteristic	FS =Functional Suitability, PE = Performance efficiency, C = Compatibility, U = Usability, R = Reliability, S = Security, M = Maintainability, P = Portability.
Attribute	Name of the attribute
Metric	Name of the metric
Function	Calculus function
Result	Possible results
Description	Brief descrption
Metric type	B = Base, D=Derived.
Tool support	M=Manual, A=Automatic.
Measurement result	Q=Qualitative, C=Quantitative, H=Hybrid.
Lifecycle phase	Req. = Requirements, Acq. =Acquisition, Dev. = Development, Int. = Integration, Ope. = Operation, Ret. = Retirement
Artifact evaluated	Spe. = Specification of cloud service, Arc. = Architecture of cloud service, Ser. = Cloud Service
Service type	SaaS= Software as a service, PaaS = Platform as a service, IaaS= Infrastructure as a service
Stakeholder' viewpoint	CSP = Cloud Service Provider, CSC = Cloud Service Consumer, CSB = Cloud Service Broker, CSD = Cloud Service Developer, USE= End-user
Validation procedure	A.A. = Axiomatic approach, A.T.M = Approach based on the Theory of Measurement, C.S. = Case study, SU. = Survey, C.E = Controlled Experiment, N.V.= Not validate, P.C. = Proof of concept

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Energy Saved	% Energy saved	% of energy saved = ((en – ep) / en)*100 en is the energy consumed in normal scenario, ep is the energy consumed by the proposed system	The range is 0...100	measures the percentage of Energy saved	D	M	C		X							X			X		X									X		S38
S	Access Control	Ac_state		access control	indicates the access control state	B	M	C					X				X			X	X											X	S48
S	Transition State	Ac_trans		state transition	indicates the state transition	B	M	C					X				X			X	X											X	S48
FS	Accuracy of service	Accuracy of service	the number (or time) that a service satisfies promised values / total number (or time) of the service in a given time interval		Positive. Accuracy is defined as the correct rate generated by the service.	D	M	C		X			X				X	X	X	X	X	X						X					S49
FS	Accuracy of service	Accuracy expected	Accuracy : (expected CS - |expected CS - observed CS|)/expected CS		measures the accuracy expected	D	M	C					X				X	X			X											X	S65
FS	Accuracy of service	Accuracy frequency	accuracy frequency = Σ fi/n, where fi is the number of times the Cloud provider fails to satisfy promised values for user i over the service time T, n is the number of previous users		the frequency of failure in fulfilling the promised SLA in terms of Compute units, network, and storage	D	M	C		X							X			X		X									X		S24

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
FS	Accuracy of service	Accuracy frequency	ϕS =Σnu/nu val, where nu is the number of users, n u-val is the amount of expected values by user u, n u-val is the amount of values actually received by user u	In the case of storage service, it is the deviation of the average read/write time. In the case of a computing service, the accuracy is defined as the difference between the achieved performance and the specified performance in the SLA. In the case of computing resources such as virtual machines, the most important accuracy indicator for a service is the number of times the SLA is violated	measures the number of times the SLA is violated	D	M	C		X							X	X			X	X									X		S30
FS	Accuracy of service	Accuracy of service	AoS = 1 - number of failed requests/total number of requests		Accuracy of services, AoS can be calculated a degree of right response for a user's request.	D	M	C					X				X	X			X										X		S53
FS	Accuracy of service	Accuracy of service	SA=(number of correct responses) /(total number of requests) where the denominator is a total number of requests from customers and the numerator is a number of correct responses without failure. The numerator can be acquired as (total number of request – number of failures).	number	measures a degree of correct response for a customer's request. The correct response means that the service provides what a service consumer wants within specific time period. The time period is defined in service level agreement (SLA)	D	M	C		X							X	X			X										X		S44

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
FS	Accuracy of service	Accuracy value	accuracy value= Σ(αt-αi)/αi Ti where α can be computational, network or storage unit of the service and Ti is service time T for user i		measures the accuracy value	D	M	C		X							X			X		X									X		S24
PE	Connections	Active connections		The number of active connections	measure the number of active connections	B	M	C					X				X			X	X	X										X	S35
PE	Turnaround Time	actual turnaround time		time in milliseconds	actual turnaround time is the exact time between the submission of a job by a user and the delivery of the completed job to the user.	B	M	C		X							X	X	X	X	X	X									X		S50
P	service Adaptability	Adaptability	Adaptability = W.CoV + W.CoM The weights for each metric which the sum is 1. where Coverage Of Variability (CoV) and Completeness of Variant Set (CoM).		Adaptability of services is the level of effectiveness in adapting services for the use of each service-based application.	D	M	C					X				X	X			X										X		S53
P	service Adaptability	Adaptability		time in milliseconds	measures the time taken to adapt to changes or upgrading the service to a higher level. Adaptability is the ability of the service provider to adjust changes in services based on customers' requests.	B	M	C		X							X			X		X									X		S24
PE	Capacity of a service	Aggregate request slowdown	∑ delays suffered by requests		measures the accumulated sum of all delays suffered by requests involved in QoS violations	D	M	C					X				X			X	X				X							X	S18
P	Agility	Agility	∑ | Cap_ prov(i) - Cap_min(i) |		measure the ability of the cloud service to scale with the increased or decreased workload	D	A	C		X			X				X			X	X	X									X		S61
PE	Agility manteined	Agility manteined	Agility manteined = ∑ Excess(i)/∑Cap_min(i) + ∑Shortage(i)/∑Cap_min(i)	this score should be as close to zero.	measures the agility manteined	D	A	C		X			X				X			X	X	X									X		S61
PE	Capacity of a service	Allocated resource-time	Allocated resource-time: resource capacity – in machine- seconds – allocated to serve user requests.		measures the allocated resource	D	M	C					X				X			X	X				X							X	S18
P	Reactivity	Allocation reactivity(AI)	AI = 1 - Σ Min(1, Ta/(Tp - Ta)) / K where Ta: allocation time, Tp: procesing time, Tr: release time	range 0..1	reflects the level of the cloud elasticity through several request. Measures the reactivity of the cloud against the workload	D	M	C		X							X			X	X				X							X	S02

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Application capacity	Application capacity		number of requests/min	measures the capacity of application by the number of request per minute	B	M	C			X		X			X	X	X			X	X		X							X		S69
S	Authenticity	Authenticity	Authenticity --> Real(0,1)	equals to 0 indicates user does not have access to the controlling privilege, while 1 means users have this privilege.	used for determining whether users have the privilege of this cloud service	D	M	C					X				X			X	X											X	S48
R	Service Availability	Availability	Availabiliy = MTTF/(MTTF +MTTR) MTTF stands for Mean Time To Failure MTTR stands for Mean Time To Recover.		the percentage of time the service is available for the end users in a given duration.	D	M	C		X							X	X	X	X	X	X			X						X		S34
R	Service Availability	Availability	Availability --> Real[0,1]	0 or 1	measures the availability of cloud service	D	M	C					X				X			X	X											X	S48
R	Service Availability	Availability	Global Provider View tool by Compuware		Compuware	D	A	C	X	X			X				X	X	X	X	X	X	X								X		S10
R	Service Availability	Availability of resource	Availability of resource Rk(AV) = Ak / Nk Nk recorded works/jobs submitted to Rk resource, Ak number of works/jobs accepted by Rk resource in T time limit		AV : Availability is the degree to which a system or component is operational and accessible when required for use, the degree to which a system or a component is achievable or usable when needed to be used	D	M	C		X			X				X	X	X	X	X	X										X	S50,S27
R	Service Availability	Availability	Availability =MTTF/MTBF where MTBF is Mean Time between Failures, MTTF is Mean Time to Failure		The ratio of service availability	D	M	C					X				X	X			X											X	S65
R	Availability Efficiency	Availability Efficiency of a SaaS provider (avEff)	avEff t-j = av SaaS t-j / (Σ k∈K t-j avCloud t-k / |K t-j|) where K t-j is the set of cloud providers that are servicing SaaS provider j at time t		defines how well a SaaS provider can provide availability to a user based on the availability it receives from the cloud provider	D	M	C	X	X							X	X			X		X								X		S45
R	Composite Service Availability	Availability of the Composite Service	Availability of the Composite Service C up to Nth minute		Availability of the Composite Service C up to Nth minute	D	A	C					X				X			X	X											X	S12

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Service Availability	Availability of web service	Availability = W.FoS + W.CoS + W.RT + W.CoM The weights for each metric which the sum is 1. FoS: Flexibility of service, CoS: Scalability of service, RT: Response time, CoM: Completeness of variant .	The range is 0 ... 1 higher value of Availability shows that the web services is highly available.	Assesses the ratio of the total time to the time which a web service is capable of being functional.	D	M	C					X				X	X			X										X		S53
R	Availability Time	Availability Time	A(t)=1 - (E[r d1-X(t)] + E[r d2-X(t)]) where rd1 checks when the drop condition is satisfied: rd1 = (P# run = M & P# queue = Q & P# send= D) ?1:0 and rd2 checks when a request is dropped due to the federated cloud unavailability: rd2 = (P# run = M & P# queue = Q & P# send <D) ?1 -af :0		measures the steady-state probability that the system is able to accept a request.	D	M	C		X	X						X			X	X	X		X								X	S14
R	Service Availability	Availability utility function	UP (X)= {uXmin + (1 + uXmin) .|RXP-X / RXP-IXP| where IXP ≤ X ≤ RXP, otherwise: 0, uX min is the minimum utility that the consumer and the provider receive for reaching a deal at its reserve performance. Qa be the availability performance quality at which a consensus is reached by both parties. Qa is substituted to X which are generalized utility functions for cloud QoS so that the performance utility is UQa(Qa) for reaching a consensus at Qa where IX=IQa,and RX=RQa.	The range is 0...100. A consumer may select availability according to their budget. Intuitively, consumers prefer the highest guarantee for availability, and providers want to sell their services with the lowest guarantee for service availability with a given price	percentage of uptime of the service in reserved timeslots.	D	M	C		X							X	X		X	X	X									X		S67

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Cost Scaling	average cost of scaling down	AVCd = P/AVTd where P: renting price per renting a min of the VM		measures the average cost of scaling down (AVCd). metric to evaluate the cloud resource elasticity	D	M	C					X				X			X	X	X										X	S35
P	Cost Scaling	average cost of scaling up	AVCu = P/AVTu wher P: renting price per renting a min of the VM		measures the average cost of scaling up. metric to evaluate the cloud resource elasticity	D	M	C					X				X			X	X	X										X	S35
PE	Latency	Average delay for a logical I/O request	Average delay for a logical I/O request.		measures the time that elapses on average between when a request is issued and when the necessary block is returned constitutes the average delay for a logical I/O request.	D	M	C					X				X		X			X			X							X	S20
P	Scaling Down	average delay time at scaling down	AVTd = ∑ scaling down delay time/number of scaling down		measures service aspect of elasticity from the delay time and the number of continuous scaling down	D	M	C					X				X			X	X	X										X	S35
P	Scaling Up	average delay time at scaling up	AVTu= ∑ scaling up delay time/number of scaling up		measures service aspect of elasticity from the delay time and the number of continuous scaling up	D	M	C					X				X			X	X	X										X	S35
PE	Energy Efficiency	Average Energy Consumption per Request (AECpR)	AECpR= total energy consumption/total successful requests been served	Energy Consumption per Request	measures the average energy consumption per request	D	M	C					X				X	X			X	X										X	S31
PE	Performance of the Cloud patterns	Average number of queries	Average number of queries		average and maximum number of queries executed by the application during one second.	D	M	C			X						X	X						X				X					S33
P	Overprovisioning	Average Overprovisioning resource	AVEs = The Dynamism can be evaluated as the average of free CPU capacity of all running hosts.		measures provisioning resource is over the needs, the SLA and QoS are not optima	D	M	C					X				X			X	X	X										X	S35
PE	Response Time	Average Response Time	Tavg-rep=∑Trepi/n , where Trepi is the elapsed time between the request sending time and the response reception time, n is the total number of transmitted requests	time in milliseconds A low value of a response time is a desirable expected from a cloud-based service	measures the average response time	D	M	C	X	X			X				X	X		X	X	X										X	,S24,S3
PE	Response Time	Average Service Response Time	ASRT ri = ∑ ri SRT RI /n where SRTRi is the time between the moment a request Ri is ready to run and the service executes the request effectively	Time in milliseconds	measures the Average Service Response Time of a request Ri (ASRTRi) executed by n physical/virtual machines	D	A	C		X							X	X			X	X										X	S17
P	Installability	Average time to install		time	average time experienced by the previous users of the Cloud service to install it.	B	M	C		X							X			X		X			X						X		S24

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
U	Learnability	Average time to learn		time	average time experienced by the previous users of the Cloud service to learn it.	B	M	C		X							X			X		X			X						X		S24
U	Operability	Average time to operate		time	average time experienced by the previous users of the Cloud service to operate	B	M	C		X							X			X		X			X						X		S24
U	Understandability	Average time to understand		time	average time experienced by the previous users of the Cloud service to understand it.	B	M	C		X							X			X		X									X		S24
P	Underprovisioning	Average Underprovisioning resource	AVEd = Underprovisioning resource/Number or Underprovisioning		measures provisioning resource is under the needs, the SLA and QoS are not optima	D	M	C					X				X			X	X	X										X	S35
M	Publicity	Awarability of Service(AoS)	Awarability of Service(AoS) AoS = 1 - No. of users with services undiscoverability / total number of fields		AoS shows that the service is more easily available to service users.	D	M	C					X				X	X			X										X		S53
PE	Bandwidth	Bandwidth		kbps, B, MBs/s , Bits/second (B/s),Gbps	network capacity metric. Binary rate of video, Data transfer rate or I/O processing speed	B	M	C	X	X	X		X			X	X	X	X	X	X	X	X	X				X				X	S69,S65
PE	Bandwidth	Bandwidth availability	|blast−bk/blast| > thr blast is the last updated value of the available bandwidth metric and bk is the current state of the metric.		measures the bandwidth availability	D	A	C					X				X	X				X										X	S05
PE	Bandwidth	Bandwidth blocking rate	bandwidth blocking rate = Σ Bandwidth of rejected requests/ Σ Bandwidth of arrived requests,		Rate between rejected request and arrived requests. An arrived request reflects the success of the algorithm to set an end- to-end QoS path according to the bandwidth metric. A rejected request indicates that the end-to-end QoS path contains at least one link that does not satisfy the bandwidth constraint.	D	A	C					X				X	X				X										X	S05
PE	Bandwidth	Bandwidth capacity		kbps, B, MBs/s , Bits/second (B/s),Gbps	measure the limits of available Amazon network's bandwidth.	B	M	C					X			X	X			X	X	X									X		S16,S10
PE	Energy Consumption	Base Energy Consumption	EBase = ttotal · P * , where ttotal is the total open time of the server S, P* is the base power of S	Energy Consumption per time	measures the base energy consumption	D	M	C					X				X	X							X							X	S19
R	Bigdata Analytics	Bigdata Analytics	An		The ability to uncover hidden information and predict the future	D	M	C					X				X	X			X	X						X					S36
PE	Bytes read from hard disk	blkio_io_bytes_read		bytes	measures the bytes read from hard disk	B	A	C					X			X	X		X		X											X	S70
PE	Bytes written to hard disk	blkio_io_bytes_write		bytes	measures the bytes written to hard disk	B	A	C					X			X	X		X		X											X	S70

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Latency	Buffer Access Latency (BAL)			Buffer Access Latency (BAL)	D	M	C					X				X		X			X			X							X	S20
PE	Capacity of a service	Capacity Avg	Capacity Avg = average utilization over a period of time of the workloads handled by the system.		The average capacity is calculated by following the average utilization over a period of time of the workloads handled by the system.	D	M	C					X				X			X	X	X										X	S58
PE	Capacity of a service	Capacity Max	capacity ξ = upper limit of the number of requests (per second) a service is able to process.	request/s	The capacity measures the upper limit of the number of requests (per second) a service is able to process. In terms of computing, the capacity depends on the virtual machine performance being used by this service and the physical machines(cluster component) hosting these VMs.	D	M	C		X							X	X			X	X									X		S30
PE	Capacity of a service	Capacity Min	Cap_min= difference between the minimum capacity needed to meet the QOS at a given workload level for an interval i	Capacity minimum	This metric measures the minimum capacity by the difference between the minimum capacity needed to meet the QOS at a given workload level for an interval i	D	A	C		X			X				X			X	X	X									X		S61
PE	Capacity of service provider	Capacity of service provider	capacityi= Σ(Fij*Wj) F=[cpu, mem, sto, bw] Wj = weights of capacity of cpu, mem, sto, bw		measures the capacity of service provider in cpu, memory, storage, bandwidth	D	M	C	X	X							X			X	X	X	X								X		S09
PE	Capacity provisioned	Capacity provisioned	Cap_prov = recorded capacity provisioned for interval i		measures the recorded capacity provisioned for interval i	D	A	C		X			X				X			X	X	X									X		S61
PE	Capacity ratio	Capacity ratio	C(p) = X (p) /X (1) The relative capacity curve C (p) is a good indicator for how much the observed behavior diverges from the ideal linear scalability C L (p) = p.		measures the ratio of the throughput of the system for a load of p , compared to the baseline of its throughput for a load of 1.	D	M	C			X		X			X	X	X			X	X										X	S55
S	Certification	Certification		Types of certification facility	Type of cerfication	B	M	Q	X	X			X		X		X	X	X	X	X	X	X	X							X		S10

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure								Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Energy Consumption	Cloud Energy consumption	$ECloud = \sum ES + \sum ESCom$ ECloud is the total energy consumption of the cloud infrastructure, ES is the energy consumption of server S, and $\sum ES$ is the energy consumption of all servers in use, ESCom is the energy consumption for communication between two servers, and $\sum ESCom$ is the energy consumption for communication among all servers.	Energy consumption	measures the Cloud Energy consumption	D	M	C					X				X	X							X							X	S19
PE	Energy Consumption	Cloud Energy consumption	$Etotal = Efix + \sum ECtsi + \sum ECtci + \sum ECtti + Esche$ where i= 1 to n Total energy consumption in the cloud environment can be summed up as the energy consumed in all three tasks namely data intensive ECtsi, computation intensive ECtci and communication intensive ECtti where tsi,tci and tti is the ID of each tasks subsequently	Energy consumption	measures the total energy consumption	D	M	C		X							X			X		X									X		S38
PE	Energy Consumption	Cloud Energy consumption	Total Energy consumption = (Fixed energy consumption+ Storage energy consumption + Computation energy consumption + Communication Energy consumption)	Energy consumption	measures the total energy consumption of the resources	D	M	C		X							X			X		X									X		S38
PE	Cloud Energy Consumption	Cloud power consumption	$E = \sum EDC + \sum ET$ E is the total energy consumption, EDC is the data centers energy consumption, ET is the transmission energy consumption.	Energy consumption	measures the total cloud power consumption	D	M	C		X							X			X		X									X		S38

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Energy Efficiency	Communication Resource Utilization	RU = fcmi(Ncmi, DPcmi, DTcmi, SCi) where Ncmi is the Process Number communication, Dpcmi is the Size of Data Processed communication, Dtcmi is the Size of Data Transmitted communication		measures the resource utilization of each task, can be calculated from the communication task	D	M	C					X				X			X	X	X										X	S01
FS	Completeness	Completeness	Completeness : total existing CSs/total requested CSs		measures the total completeness	D	M	C					X				X	X			X											X	S65
R	Completeness of variant set	Completeness of variant set(CoM)	CoM = Σ (number of variants supported in variant point/total number or potencial variants in variant point) /n	The tendency of the throughput is high, meaning the larger the value the better the service performance is	how many choices are rendered by services so that service users can opt. one of them for their need	D	M	C					X				X	X			X										X		S53
PE	Completion time	Completion mean time	Mean time of response: avg (time of response)	time (sec.)	measures the mean time of response for asynchronous task	D	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Completion time	Completion standard deviation time (asynchronous task)	Standard deviation time of response	time in milliseconds	measures the standard deviation time of response for asynchronous task	D	M	C			X		X			X	X			X	X	X		X							X		S69
C	Composability	Composability	Composability = W.SM + W.SI Service Modularity, SM, and Service Interoperability SI, W weight assigned based on the importance The weights for each metric which the sum is 1.	The range of Composability is 0 ... 1 and higher value of Composability	Composability indicates that the web services by incorporating other services in with is more easily and efficaciously customized to service users-specific need	D	M	C					X				X	X			X										X		S53
PE	On Going Cost	Compute usage		cpu units	measures the compute usage	B	M	C		X							X			X		X									X		S24
PE	Computing Capacity	Computing Capacity	Computing Capacity = Actual Usage time of the Resource/Expected Usage time of the Resource		measures the computing capacity of the resource	D	M	C					X				X		X		X											X	S65
PE	Conference start time	Conference start time	Conference start time= time conference ready - time receipt of a request	time	measures the time required to get a conference ready upon the receipt of a request	D	M	C					X			X	X	X			X											X	S04
PE	Network capacity	Connections utilization	Connections utilization = current connections / total available connections *100	ratio, The range is 0...100.	measures the percentage of current connections in total available connections	D	M	C			X		X				X			X	X	X										X	S35

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
FS	Consistency	Consistency Across Several Instances	number of deviation from the promised values over a set of similar instances. Promised values of CPU speed, Memory, etc.	number of failures	measures any deviation from the promised values of CPU speed, Memory, etc. over a set of similar instances. Any deviation is considered a failure.	D	M	C	X	X							X			X		X	X								X		S71
FS	Consistency	Consistency Over a Period of Time	Number of changes in the instance parameters	number of failures	measures any deviation from the promised values of CPU speed, memory etc. as a cloud instance ages over a period of time. Any change in the instance parameters is considered a failure	D	M	C	X	X							X			X		X	X								X		S71
PE	Storage utilization	Content_cc	Content_cc= (Content Data_scale --> Integritymetrics)		used to compare the size of data with the storage capacity	D	M	C					X				X			X	X											X	S48
PE	correlation performance security	correlation performance security	E(δ(x)) = Σ Na(x, n) · δ(x) (Na(x, n)) reflects the impact of security factors on the number of available VMs. (δ(x)) reflects the impact of the number of available VMs on the service performance.		represents the new performance metric taking the impact of security into consideration.	D	M	C					X				X			X	X				X							X	S79
P	Cost Scaling	Cost (c)	c = (AVCu +AVCd) /2 average price AVCu and AVCd to describe the cost of scaling up and scaling down		measures average price to describe the cost of scaling up and scaling down	D	M	C					X				X			X	X	X										X	S35
R	Failure Recovery	Coverage of Failure Recovery (CFR)	CFR=(number of failures remedied) /(total number of failures), where the denominator is the total number of failures caused by faults	number	measures the ratio of remedying failures the specific period of time.	D	M	C		X							X	X			X										X		S44
R	Fault tolerance	Coverage of Fault Tolerance (CFT)	CFT =(number of faults without becoming failures)/(total number of faults occurred), where the denominator is the total number of occurring faults identified and the numerator is the number of faults which do not cause failures	number	measures the ratio of bearing the occurrence of a fault without a failure.	D	M	C		X							X	X			X										X		S44

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Coverage of Scalability	Coverage of Scalability	COS = ∑(amount of allocated resources of ist request)/(total amount of requested resources of ist request) /k where k is the total number of requests for extending resources used	The range is 0…1. The value 1 implies that all requested resources are allocated. A SaaS with a higher COS value yields a higher acceptance for the customer's request to re-scale the resources used.	measures the average amount of allocated resources among the amount of request resources. reflects the ability of the cloud to fit the increase of the workload	D	M	C		X			X				X	X		X	X				X						X		S26,S02,S44,S
M	Variability	Coverage of Variability	CoV = 1 - No.variation point not supported / total number of Variation points , where denominator the total number of potential variation points by considering and protruding its set of users applications, numerator is the number of variation points not supported by the service		Coverage of Variability: CoV is measured by considering how many variation points can be customized by service users	D	M	C					X				X	X			X										X		S53
M	Variability	Coverage of Variability rate	CV = (number of variation points realized in the SaaS) / (number of variation points in the domain), where denominator is the number of variation points identified for SRSs in the domain.	The range is 0...1. A higher value indicates that the coverage of the SaaS is comprehensive enough to have realized a larger number of variation points.	Coverage of Variability: This metric measures how many of the variation points included in the domain are actually realized in the SaaS.	D	M	C		X							X	X			X										X		S44
PE	CPU capacity	CPU capacity	CPU capacity = number of CPU * frequency of CPU		measures CPU capacity using the number of CPU and frequency of CPU	D	M	C	X	X			X				X			X	X	X	X								X		S09,S10
PE	CPU capacity	CPU Frequency		Mhz	measures the CPU Frequency	B	M	C			X		X			X	X			X	X	X		X							X		S69
PE	CPU Downtime	CPU performance		number of iterations/second.	measures the number of iterations per second.	B	A	C		X							X			X		X										X	S81

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	CPU Downtime	CPU Performance Coefficient(CPU-PC)	CPU-PC = cpu unavailable/ cpu cores * ti	time in milliseconds. Closer to 1 represent better CPU performance, closer to 0 represent worse CPU performance	measures the CPU Performance	D	M	C					X				X			X	X	X										X	S51
PE	Energy Efficiency	CPU Power consumption	Power consumed by the CPU = k * Pmax + (1- k) * Pmax * u where Pmax is the maximum CPU utilization	Energy consumption	measures the total server power consumption against the CPU utilization	D	M	C		X							X			X		X									X		S38
PE	CPU utilization	CPU utilization	CPU utilization = current occupied computing resource / total computing resource	ratio, The range is 0...100.	measures the percentage of current occupied computing resource in total computing resource	D	M	C					X				X			X	X	X										X	S35
R	CPU Downtime	CPU_unavailable	Actual unavailability time of the CPU(CPU_unavailable)	time in milliseconds	Actual unavailability time of the CPU. Amount of time in milliseconds	D	M	C					X				X			X	X	X										X	S51
PE	CPU usage Container	cpu_usage		ratio, The range is 0...100.	measures the percentage of CPU usage of container	D	A	C					X			X	X		X		X											X	S70
PE	CPU usage VM	cpuUsedPercent		ratio, The range is 0...100.	measures the percentage of CPU utilization of VM	D	A	C					X			X	X			X	X											X	S70
PE	Reset state	CReset		CReset: Trans -->R ©	measures the reseting clock after state transition	B	M	C					X				X			X	X											X	S48
M	Customizability	Customizability	Nd/(Nd+Ns) where Nd: Number of dynamic changes in a Cloud service with respect to a workload. Ns: Number of static changes in a Cloud service with respect to a workload.		measures the customizability degree	D	M	C					X				X	X			X											X	S65
PE	Data centre location	Data centre location	yi≥x, yi = number of data centres provided by provider, x = number of data centres required by an application	Providers who does not satisfy constraints , will be sorted by their number of data centres and ranked	measure number of data centres their interdistance and the area where a user wants the data centre services.	D	M	C	X	X			X		X		X	X	X	X	X	X	X								X		S09,S10
PE	Data communication	Data communication		data units	measures the data communication	B	M	C		X							X			X		X									X		S24

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
S	Data Integrity	Data Integrity	Data Integrity of resource Rk(DI) = Dk / Ck Ck jobs completed successfully by resource Rk, Dk denote the number of jobs data integrity preserved by resource Rk over the period T, DI= Dk / Ck Ck number of works completed by Rk source Dk number of data integrity works provided by Rk in T time limit, checks if the resource correctly stores data		Data integrity is a broad term and it includes security, privacy and accuracy of the data., is a broad term and includes: security, latency and accuracy of the data,	D	M	C		X			X				X	X	X	X	X	X										X	,S27,S3
R	Data loss	Data loss	MTTDL mean time to data loss	time	mean time to data loss	D	A	C		X			X				X			X	X	X									X		S61
PE	Data transfer	Data transfer rate			measures how fast the service is provided	D	M	C	X	X			X				X			X	X	X		X							X		S26
PE	Data transfer	Data Transfer Time	Dj = ∑ DT i,l Dj is the maximum data to be stored in the storage resource Dsj associated to VMI where DT i,l ≠ 0 if the task i and l are related in the Workflow		measures the transferring data from one service to another using a specified network and it is calculated only on destination services. Data transfers are determined by the source services and the destination services.	D	M	C					X				X			X	X	X										X	S13
PE	Data transmission performance	Data transmission speed achieved	Mean value and standard deviation of transmission rate	kbit/s	measures the data transmission speed achieved	D	M	C					X				X	X			X										X		S63
S	Data Integrity	data_alteration_prevention	verifies if the resource correctly implements security measures to preserve data integrity.		measures the data_alteration_prevention	D	M	C					X				X	X			X	X										X	S39
S	Data Integrity	data_alteration_slo	precondition says that the SLO is valid whenever the evaluation of the composite metric is true		measures the data_alteration_slo	D	M	Q					X				X	X			X	X										X	S39
S	Replication Policy	data_replication			measures the data_replication. Verifies if the resource correctly deploys the replication policy	B	M	C					X				X	X			X	X										X	S39
R	Database Backup	Database Backup		size in GBs	measures the Database Backup for every cloud service (CS)	B	M	C					X				X	X			X											X	S65

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	DB sever	DB sever commits		number of commits/second	measures the number of commits per second	D	M	C					X				X		X		X											X	S65
R	Defective Operation	Defective Operations Per Millon attemps (DPM)	DPM= (OperationsAttempted - OperationsSuccessful) / OperationsAttempted * 10^6		the ability of an item to perform a required function under stated conditions for a stated time period	D	A	C					X				X	X			X	X			X				X				S03
R	Defective Operation	Defective Operations Per Millon attemps (DPM)	DPM= OperationsFailed / OperationsAttempted * 10^6		the ability of an item to perform a required function under stated conditions for a stated time period	D	A	C					X				X	X			X	X			X				X				S03
R	Defective Operation	Defective Operations Per Millon attemps (DPM)	DPM= OperationsFailed / (OperationsSuccesful+ OperationsFailed) * 10^6		the ability of an item to perform a required function under stated conditions for a stated time period	D	A	C					X				X	X			X	X			X				X				S03
M	Defects	Defects	Number of Defects reported= Defects/CS	Number of Defects	measure the number of defects per cloud service	D	M	C					X				X	X			X											X	S65
R	Resiliency	Degradation time	Tα,deg = td - tb It is the time Tdeg elapsed between the burst begin (tb) and the instant (td) at which the performance index reaches a value lower to the α- 100% of its steady-state value		measures the degradation time. It can be considered as an estimation of the time needed to observe the effects of the burst on the selected performance index.	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Delay	Delay		time in milliseconds	The core network that will be used only as a transport network may introduce inherent propagation delays., Mean value and standard deviation of the delay (in milliseconds).	B	M	C			X		X			X	X	X		X	X	X		X							X		S69,S16,S6
PE	Delay	Delay		time in milliseconds	Measured as the round trip time for packets, simply using the standard Ping command.	B	M	C					X				X	X			X										X		S54
PE	Delay	Delay Time	T Dij= Trij + Tnij + TOg where Trij is the response time from the cloud storage, TOg is the delay time in the gateway caused by the redirection overhead, and Tnij is the network time the request takes to travel from client i to cloud j.	time in milliseconds	measures the accumulated delay from the resources, resource redirection and network.	D	M	C	X	X			X				X			X	X	X										X	11,S26
PE	Delay	Delay Time	Delay time: time getting resource - time requesting resource	time in milliseconds	Delay time: delay time is the period of time from requesting resource to getting resource	D	M	C					X				X			X	X	X										X	S35

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Delta	Delta		time in milliseconds	measures the packet inter-arrival time per RTP packet	B	M	C		X			X				X			X	X	X	X					X					S23
PE	Density	Density App		number of application servers	mesures the number of application servers running on a system	B	A	C					X				X		X		X	X									X		S61
PE	Density	Density User		number of users	measure the number of users that the system can service	B	A	C					X				X	X			X	X									X		S61
PE	Density	Density VM	number of virtual machines running on a piece of physical software	number of virtual machines	measures the number of virtual machines running on a piece of physical software. How many workload instances can be run on the cloud system at once before the performance degrades below a specified QoS	D	A	C					X				X			X	X	X									X		S61
P	Dependability	Dependability	DS(e,G), is a measure of how good the system S meets its goal G under the environment condition e ∈ E∗.		Dependability as application-level QoS at a meta-level (regardless of the system complexity). Dependability assessment of a cloud-based system involves three aspects: measurability, predictability, and adaptability of system behavior. • Measurability, wherein S can map its output/state vari-ables onto to high-level metrics over a wide range of input conditions (for analysis and reasoning); • Predictability, wherein S can compute its expected output under various input and environment conditions (with reasonable accuracy); • Adaptability, wherein S can adjust the utility extracted from its output as the environment conditions become harsher, thereby heavily dwindling its resources.	D	M	C			X						X	X		X				X							X		S59
P	Deployment latency	Deployment latency	deployment latency = time actual provisioning of that cloud instance - time requesting for a cloud instance	time in milliseconds	measures the time elapsed between requesting for a cloud instance and the actual provisioning of that cloud instance. A good measure is achieved while provisioning 10, 100, 1000 instances	D	M	C	X	X							X			X		X	X								X		S71
PE	Performance of Virtual Machine	Disk input output (I/O) performance		Kb/s	refers to the property of storing data on a permanent basis, and it corresponds to the evaluation of the VM Disk storage	D	A	C		X							X			X		X										X	S81
PE	Storage capacity	Disk read		Gb/s	measures the speed of reading disk	B	M	C					X				X			X	X	X										X	S35

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Disk utilization	Disk utilization	Disk utilization = current occupied disk resource/total disk resource	ratio, The range is 0...100.	measures the percentage of current occupied disk resource in total disk resource	D	M	C					X				X			X	X	X										X	S35
PE	Storage capacity	Disk write		Gb/s	measures the speed of writing disk	B	M	C					X				X			X	X	X										X	S35
PE	Disk utilization	diskFree		size in GBs	measures the amount of available disk capacity	B	A	C					X			X	X			X	X											X	S70
PE	Disk utilization	diskUsed		size in GBs	measures the amount of used disk capacity	B	A	C					X			X	X			X	X											X	S70
PE	Data centre location	Data centre distance	For finding data centres as per the requirement of application find the distance between each service location and data centre and add them.	A provider with minimum sum of distances will be ranked 1 and so on.	measure the distance between location of data centre and expected service location	D	M	H	X	X							X			X		X	X								X		S09,S10
FS	Distortion in Interval (DIV)	Distortion in Interval (DIV)	DIV=maximum percentage of frames with a PSNR worse than original	percentage	measures the maximum percentage of frames with a PSNR worse than original	D	M	C		X			X				X			X	X	X	X					X					S23
R	Downtime	Downtime	Downtime = (Uptime2.PreciseTimeStamp - Uptime1. PreciseTimeStamp) - Uptime 2		Duration of service not available	D	A	C					X				X	X			X										X		S15
R	Downtime	Downtime%	UAv(s)= (ΣSF(t) / AG(t)) x 100% SF(t) is the service unavailable time during the negotiated time, AG(t) is the negotiated time.		UAv(s) is the percentage of service unavailable time during the negotiated time.	D	A	C					X				X	X	X	X	X	X									X		S08
R	Durability	Durability	Durability = (MTTF/ (MTTDL *MTTR)) MTTF represents mean time to multiple failures, MTTDL represents mean time to data loss MTTR represents the mean time to repair (downtime)		Durability is the probability of data loss	D	A	C		X			X				X			X	X	X									X		S61
R	Downtime	Duration of a single outage	outage end time - outage begin	time in miliseconds	measures the duration of service not available	D	M	C			X		X			X	X	X		X	X	X		X							X		S69
M	Customizability	Dynamic changes	Number of dynamic changes in a Cloud service /workload (Nd)		measures the dynamic changes in a Cloud service with respect to a workload	D	M	C					X				X	X			X											X	S65

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Latent Capacity	Dynamism or “Latent Capacity”,	If the hosts have DVFS enabled, the Dynamism is computed rel- atively to the maximum CPU capacity and the current capacity used Dyn = (ξ f,cpu −ω f,cpu)/nH, where f is the highest CPU frequency allowed and nH only includes hosts that are powered On		measures Dynamism as the average of free CPU capacity of all running hosts.	D	M	C		X							X	X			X	X										X	S30
PE	Effective arrival	Effective arrival	$\eta_e = \lambda(1 - q_L)$ λ is the arrival rate of tasks. qL represents the probability that the task queue is full. When the queue is full, a new arriving task is abandoned		Effective arrival means that the task has been placed in the task queu without being abandoned.	D	M	C					X				X			X	X				X							X	S79
PE	Efficiency	Efficiency	% max Performance (speedup or utilization) achievable	ratio, The range is 0...100.	measure the rate of performance achievable	D	M	C		X			X				X		X		X	X						X					S36
PE	Efficiency	Efficiency	Efficiency= W.RU + W.TB where Time Behaviour (TB), Resource Utilization (RU)	The range is 0...1 Higher value of Efficiency indicates that the SaaS is more effective.	measures the efficiency of resources in terms of time behaviour and resource utilization	D	M	C		X			X				X	X			X										X		S44,S53
PE	Efficiency	Efficiency	System efficiency = E=T/(T+To) T: the execution of tasks of an application in traditional data center To: the overhead of Cloud data centre	A higher value for efficiency indicate that the overhead will be smaller	The Cloud system efficiency indicates the effective utilization of leased services.	D	M	C	X	X			X				X			X	X	X	X								X		S24,S10,S0

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Cloud elasticity	Elasticity	Ec = Tj/ Tm = 1 −(Tu +To)/Tm percentage of time when the platform is in just-in-need states. Tu is the under-provisioning time that the cloud platform needs to switch from an under-provisioning state to a corresponding balanced state To denotes the over-provisioning time in which cloud platform needs to switch from an over-provisioning state to a balanced one in those periods of time.	percentage	measures the ability to expand and contract overtime in response to user demands. It is the power to scale computing resources up and down easily when responding to those demands.	D	M	C					X			X	X			X	X	X									X		S26
P	Cloud elasticity	Elasticity	Eq =1/6 (AVEs + AVEd) (AVTu + AVTd) (AVCu + AVCd) where average cost of scaling down (AVCd), average cost of scaling up (AVCu), average delay time at scaling down (AVTd), average delay time at scaling up (AVTu), Average Overprovisioning resource (AVEs) Average Underprovisioning resource (AVEd)		measures the larger is the volume of the polyhedron, and the lower is the performance of elasticity.	D	M	C					X				X			X	X	X										X	S35

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Cloud elasticity	Elasticity	Ra= the capacity variation to the previous capacity/ the necessary time for this reconfiguration(Trcf sc∗,∗). Then Ra is defined by two variables: the necessary time Trcf sc∗,∗ to increase or decrease the service and the maximum capacity variation of the service. The maximum processing capacity of a service is considered as the number of computing units and their respective maximum capacity that can be allocated during an attendance peak.		Ra is the capability of a service to “scale up” during an attendance peak. The Reactivity can also be considered as the system “speed” to increase its performance level while remaining within its acceptable scalability zone. the reactivity of a service s is estimated by the ratio of the capacity variation to the previous capacity, divided by the necessary time for this reconfiguration	D	M	C		X							X	X			X	X									X		S30
P	Cloud elasticity	Elasticity	El,= 1/(θ*μ) where θ= the average time to switch from an under-provisioned state to an elevated state, μ= be the offset between actual scaling and the auto scaling.		Dynamic interval of auto-scaling resources with workload variation in order to increase the elasticity of a cloud system, we should minimize the provisioning time and keep the provision offset as low as possible.	D	A	C					X				X		X	X	X	X	X					X					S36
P	Cloud elasticity	Elasticity	ELAS= ε = Σ r i1 / Σr i2 where r i1 y r i2 denote the i1, i2 amount of resources allocated and requested in the ith request, respectively, and n is the number of requests that issued in an operational period.	The range is 0..1 closer the value to 1, the higher the elasticity	refers to the ability of Cloud services to provide resources on demand during a time interval (e.g. storage space on demand in a time interval).	D	M	C					X				X	X	X	X		X						X	X			X	S82,S83
P	Cloud elasticity	Elasticity rate	Elasticity=w1*time+w2* cost		Elasticity can be understood by its two fundamental elements: Time and Cost. Time means how much time a service provider takes to provision or de-provision resources and cost means whether service provider charges on per hour basis or per minute basis.	D	M	C	X	X			X				X			X	X	X	X								X		S09,S10

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Performance of matching algorithms	Ending Time(Te)	TE = te − tl	Time in milliseconds	ending time TE as the time spent on additional processing after the last matched subscription is found	D	M	C					X				X	X							X							X	S56
PE	Energy Consumption	Energy consumption	represents the amount of energy consumption that VMj consumes to process αij data block for request Ri		measures the energy consumption	D	M	C					X				X			X	X	X									X		S75
PE	Energy Consumption	Energy Cost	P(Fi) = α (Pfull(Fi) − Pidle(Fi) + Pidle(Fi), where α is the CPU load, Pfull(Fi) and Pidle(Fi) are the power given by the host using the Fi frequency at 100% and 0% of CPU utilization.	Energy Consumption	measures the energy consumption	D	M	C		X							X	X			X	X										X	S30
PE	Energy Efficiency	Energy efficiency	EE = (∑Storage capacity + ∑ Network capacity)/Total Energy	percentage	measures the percentage of total facility power taken for storage and network capacity	D	M	C	X	X			X				X			X	X	X										X	S07
PE	Capacity of a service	Excess capacity	Excess(i) = Cap_prov(i)–Cap_min(i), when Cap_prov(i) > Cap_min(i) and zero otherwise.	0 to Capacity provisioned	measures the excess capacity for interval i	D	A	C		X			X				X			X	X	X									X		S61
PE	Performance of the infrastructure	Execution Cost	C = (Ci * T) + (Cs * S * T) + (Cb * B) Where C = Total Cost ci = Cost of the cloud instance per hour cs = Cost of the storage per Megabits per hour (In most cases this is bundled with the instance cost and so is zero.) cb = Cost of the data transfer per Kilobits T = Time taken for the task to run rounded off the nearest higher whole number. S = Storage consumed in Megabits B = Data transferred in Kilobits		measures the cost of executing any task on a cloud service	D	M	C	X	X							X			X		X	X								X		S71

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: Theoretical Validation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Execution time	Execution Time	Execution Time=Treq,s ex= ξ req,s f,cpu / ωv,h.∗,∗ Execution Time depends on the complexity of the request req to be executed (i.e. number of instructions, denoted ξreq,s-f,cpu) and the VM capacity, denoted ωv,h∗,∗ , where the request is processed.	time (in seconds)	measures the execution Time of VMs, knowing the MIPS capacity of these VMs and the number of instructions (in Million of Instructions) that VMs have to execute. This metric has to be minimized to minimize the Response Time.	D	M	C		X							X	X			X	X									X		S30
PE	Execution time	Execution time		∑TPS time (sec., hours)	measures the time elapsed during program or job execution, (sec., hours), ,	B	M	C		X			X		X		X	X	X	X	X	X			X			X				X	,S13,S1
PE	Response Time	Expected Request completion time	E(RCT) = E(FLT) + E(RTq) FLT denote the interval between the time when the first job arrives at PM queue and the time when the last one g job is done. RTq Resonse Time for queue		the interval between the time that a request arrives at the CMU queue and the time that its last job is done.The completion time can be divided into CMU handling time and execution time	D	A	C					X				X			X					X						X		S76
R	Failure Recovery	Failure Recovery	FR = 1 - number of failure not rectified/total number of failure		Failure Recovery (FR) metric measures the ratio of rectifying failures at a specific period of time.	D	M	C					X				X	X			X										X		S53
R	Fault Tolerance ratio	Fault Tolerance ratio	FT = 1 - number of faults becoming failure/total number of faults occured		Fault Tolerance (FT) metric measures the ratio of enduring the happening of a fault without a failure.	D	M	C					X				X	X			X										X		S53
R	Fault tolerance	Fault tolerance(FTs)	FTs = f(τM, τN, τSW) If it is assumed that the physical machine and network errors make the service unreachable, then only software errors (τSW) should be taken into account.		FTs, represents its capability to remain reachable and working when anomalies occur. Anomalies are subject to: - The error rate of a physical machine, denoted τM - The network error rate, denoted τN - The software error rate: invalid, incomplete or contradictory inputs, denoted τSW	D	M	C		X							X	X			X	X									X		S30
PE	Data Transfer	File transfer rate		MB/second	measures the file server transfer in MB per second	B	M	C					X				X		X		X											X	S65

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Fineness of Flow Control	Fineness of multiplexing(F)	f ∝ p'/p where f is the fineness of multiplexing, which is the ratio of p' and p (optimal impact power and the impact power)	ratio, The range is 0...100.	the fineness of multiplexing is proportional to the ratio of the optimal impact power and the impact power	D	M	C					X				X			X					X							X	S80
M	Flexibility	Flexibility	upgrading the service/ total time		the capability of the service provider to regulate the changes in services according to the customers' requests.	D	M	C	X	X			X				X			X	X	X										X	S07
M	Flexibility	Flexibility	rating of the ability to add or remove predefined features from a service in order to accommodate users' preferences	rating	This is the rating of the ability to add or remove predefined features from a service in order to accommodate users' preferences	D	M	Q		X			X				X	X			X	X										X	S21
R	Flexibility of service	Flexibility of service(FOS)	FOS= 1 - unavailable time for invoking web service/total timefor opearing the web services		Flexibility of service, FOS can be measured as the ratio of unavailable time for invoking the service to the total amount of operating the services.	D	M	C					X				X	X			X										X		S53
PE	Bandwidth	Flow in		kbps, B, MBs/s , Bits/second (B/s),Gbps	measures the bandwidth of flow in, and the flow isn't limited	B	M	C					X				X			X	X	X										X	S35
PE	Bandwidth	Flow out		kbps, B, MBs/s , Bits/second (B/s),Gbps	measures the bandwidth of flow out, and the peak flow decided by ECS	B	M	C					X				X			X	X	X										X	S35
M	Functional Commonality	Functional Commonality(FC)	FC = (∑number of requirements appling ith functional feature / total number of requirements analyzed in the domain)/n, where feature is a function or characteristic in a SRS to specify functionality of SaaS, n is the total number of functional features. The number of functional features is identified from SRSs in the same domain.	The range of FC is 0...1. The value 1 denotes all features applied in the SaaS are common in the same domain.,	measures an average of commonality of each functional feature defined in a target SaaS.	D	M	C		X			X		X		X	X			X										X		S53,S44

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Response Time	Gaming Response Time	RT= Rdelay + Ds + Dr + Df + De + Dc Where Df = 1/(2*FR) Ds as service delay, Dr as rendering delay, Df as average value of Tframe, De encodinf delay, Dc coding delay	Time in milliseconds	Indicate the gaming response experience	D	A	C					X				X	X							X							X	S73
PE	Performance of matching algorithms	Gap Time(Tg)	TG = TDl −TDf where TDl and TDf are the determining time of the last and the first matched subscription	Time in milliseconds	TG represents how much more time is taken to determine the last matched subscription than the first one	D	M	C					X				X	X							X							X	S56
PE	Group of Pictures (GOP)	Group of Pictures (GOP)		[size]	measures the Group of Pictures or Frames with inter dependences for decoding	D	M	C		X			X				X			X	X	X	X					X					S23
PE	Energy Consumption	Host Energy consumption	P=I*V I: corriente V: voltaje	Energy consumption	measures the host total energy consumed	D	M	C					X				X			X	X	X										X	S01
PE	Energy Efficiency	IaaS Energy efficiency	Ef Infrastructure = Σi Ef Nodej / n	Energy consumption	measures the energy efficiency of the whole infrastructure for IaaS providers. Can be evaluated as the mean energy efficiency of all the nodes comprising it	D	M	C					X				X			X		X										X	S41
R	I-Frame Loss Rate	I-Frame Loss Rate	IFLR= number of I-frames lost/ number of I-frame sent	The range is 0...100.	I-Frame Loss Rate between number of I-frames lost and sent	D	M	C		X			X				X			X	X	X	X					X					S23
PE	Fineness of Flow Control	Impact energy (E)	e ∝ l2 where e represents impact energy and l denotes the length/amplitude of the continuous content.		measures the impact energy to downstream devices is proportional to the square of the amplitude/length of the continuous content,	D	M	C					X				X			X					X							X	S80
PE	Fineness of Flow Control	Impact power(P)	p ∝ e/t (15) where t is the time span of exerting the impact energy e and p denotes the impact power.	ratio of impact energy and time	measures the impact power to downstream devices is proportional to the ratio of impact energy and time	D	M	C					X				X			X					X							X	S80
PE	Connections	Inactive connections		The number of inactive connections	measures the inactive connections	B	M	C					X				X			X	X	X										X	S35
R	Continuity	Incidence of accidents (IoA)	when a < b, x = 1-a / b; when a >= b, x = 0; x = IoA, 0≤x≤1; a = the number of accidents; b = the number of allowed accidents in the service agreement	The range is 0...1 The closer the value of “x” is to 1, the better the IoA.		D	M	C					X				X	X	X	X		X										X	S84
PE	Performance for video streaming	Initial Buffering Time (IBT)	duration = time beginning of the playback - starting time of loading a video	Time in milliseconds	measures the duration separating the starting time of loading a video and the beginning of the playback	D	M	C					X				X	X							X						X		S64

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Instance mean starting time	Instance mean starting time	Avg (Time up - time requested)	Time in minutes	measures the average between the difference of time up and time requested	D	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Instance standard deviation starting time	Instance standard deviation starting time	Standard deviation of (Time up - time requested)		measures the standard deviation between the time up and time requested	D	M	C			X		X			X	X			X	X	X		X							X		S69
R	Service Availability	Instant service probability at time t (I(t))	I(t)= E [ri X(t)] Instant service probability at time t		Instant service probability at time t. It is the probability I(t) that a request is immediately served and it corresponds to the probability that the system has at least an idle resource at time t.	D	M	C		X	X						X			X	X	X		X								X	S14
R	Instantaneous Availability	Instantaneous Availability	A(t) = R(t) + R(t − u)m(u)du where m(u) is the system renewal density function		Availability is defined as the probability that the system is operating properly at a given time t and when it is accessed for use.	D	M	C					X				X			X	X	X									X		S52
S	Integrity	Integrity	Integrity =∑ (1-Threat) x (1 -Security)		where the Threat is the probability of occurrence of an attack (of specific type) in given time and Security is the probability of repelling the specific attack in Cloud service.	D	M	C					X				X	X			X											X	S65
S	Integrity	Integrity	Integrity= (Content, Content_cc ,Integritymetrics)		describes the integrity of storage data in cloud	D	M	C					X				X			X	X											X	S48
S	Data Integrity	Integritymetrics	Integritymetrics--> Real ≥ 0	The range is 0...100.	measures the percentage integrity of data	D	M	C					X				X			X	X											X	S48
PE	Accessibility	Internet Accessibility	IA = Number of request time out/Total number of requests or Percentage of Time Out Request	The range is 0...1	measures the ratio of the number of request time out to the total number of requests for a particular resource or service in the response of particular workload	D	M	C					X				X	X			X											X	S65
C	Interoperability of service	Interoperability dependency	SI = No. of depedent services with satisfactory interaction / total no. of depedent services in participating BP's, numerator is the number of dependent services with satisfactory interaction capacity, denominator is the total number services with dependency on the present service		measured by considering the level of efficient interactions between the present web service and its subordinate services in all the BP's which the present web service participates	D	M	C					X				X	X			X										X		S53
C	Interoperability of service	Interoperability platforms	number of platforms offered by the provider / number of platforms required by users for interoperability		measures the ratio of Interoperability platforms provided	D	M	C		X							X			X		X									X		S24

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure								Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
C	Interoperability of service	Interoperability resources	the number of resources offered by the provider / the number of resources required by users for interoperability	Positive	measures services interoperability between different development environments used to implement services.	D	M	C		X			X				X	X	X	X	X	X						X					S49
PE	Performance of matching algorithms	Interval Time (Ti)	Ii,j = TDj −TDi		We define Ii,j as the interval between the determining times of the ith and jth matched subscription	D	M	C					X				X	X							X							X	S56
PE	Video Quality	Intra-coded Block Size		If the scene is simple, the intra-coded block size will be small. If the scene is complicated, the IBS will be large to contain all the information.	IBS measures the spatial aspect of video (scene complexity).	B	A	C					X				X	X							X							X	S66
PE	Video Quality	Intra-coded Macroblocks		A video with visual changes from frame to frame will have these changes encoded (either by neighbouring blocks or independently of other blocks), while video without visual changes can skip much of the encoding (PFIM).	PFIM measures the temporal aspect of the video(motion in subsequent images)	B	A	C					X				X	X							X							X	S66
PE	IOPS	IOPS		IO request/second	measures the IO request per second	B	M	C	X				X				X			X	X	X										X	S35
PE	Java order rate	Java order rate		number of new orders/second	measures the number of new orders per second	B	M	C					X				X		X		X											X	S65
PE	Jitter	Jitter		Time in milliseconds lower than 20 ms in VoIP applications	measures the variation in the time between packet arrivals caused by network congestion, timing drift, or route change.	B	M	C					X			X				X	X										X		S16

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Jitter	Jitter	J=J'+(|D(i-1,i)|-J')/16. J is jitter calculated continuously every time a Ping packet is received, J' is the previous jitter value, and t |D(i,j)| which is the value of the difference in Ping times between the i'th and j'th packets.	Time in milliseconds	Measured based on variation in delay. The jitter is calculated continuously every time a Ping packet is received	D	M	C					X				X	X			X										X		S54
PE	Jitter	Jitter		Time in milliseconds	measures the delay variation	B	M	C			X		X				X			X	X	X	X	X				X			X		S23,S69
PE	Latency	Latency	Tnet(n) = l + ß(n), where l is a constant (i.e. Latency) and ß(n) the time needed to send n bytes.		Latency represents the necessary time for a user's request to be handled by the relevant service. This time only depends on the performance and the network conditions when the user's request is sent to the service provider.	D	M	C		X							X	X			X	X									X		S30
PE	Latency	Latency	Latency =Time of input a Cloud workload - Time of output produced with respect to that Cloud workload		measures between the Time of output produced with respect to that Cloud workload and Time of input a Cloud workload	D	M	C					X				X	X			X											X	S65
PE	Latency	Latency	time = time arrival at its destination - time submitting a packet	Time in milliseconds	measures the difference between the moment when a first packet bit passed the input checkpoint and that when the last packet bit passed the output checkpoint.	D	M,A	C					X				X		X	X	X	X	X					X				X	S36,S61
PE	Latency	Latency	Avg (RequestTime-ResponseTime)	Time in milliseconds	Latency=the time that has elapsed between a request and the corresponding response	D	A	C					X				X	X			X	X			X				X				S03
PE	Latency	Latency Read		Time in milliseconds	measures the Latency Read	D	A	C					X				X	X			X											X	S37
PE	Latency	Latency Update		Time in milliseconds	measures the Latency Update	D	A	C					X				X	X			X											X	S37
U	Learnability	Learnability	Learnability : 1/ (Time taken to learn the service)		measures the time taken to learn the service	D	M	C					X				X	X			X											X	S65
PE	Load average	Load average	avg (number of load)	number of load	measures the average number of load	D	M	C					X				X			X	X	X										X	S35
PE	Location	Location			location affinity	B	M	C	X	X		X			X		X	X	X	X	X	X	X								X		S26
PE	Mail action rate	Mail action rate		number of actions/minute	measures the number of actions per minute	B	M	C					X				X		X		X											X	S65
PE	Execution time	Main Loop Iteration Time (MLIT)	MLIT= tSceneDataUpdated – tSimulatePhysics	Time in milliseconds	measures the time spent to perform one iteration of our workflow	D	M	C			X		X			X		X	X	X	X			X							X		S32

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
M	Maintenance reliability	Maintenance reliability	maintenance cost is calculated in terms of the amount of capacity that each edge/node needs to be restored. The total costmto recover the edges/nodes in a CCN from the state X is TC (X) = Σci(Wi - xi) where ci(Wi - xi) is the maintenance cost for ei on any MP to recover from the current capacity xi to its highest capacity Wi.		total maintenance cost can not exceed the budget B	D	M	C					X				X			X	X											X	S46
PE	Cloud patterns	Max Queries processed	Max (queries processed per second)	The number of queries	measures the maximum queries processed per second	D	M	C			X						X	X						X				X					S33
PE	Capacity of a service	Capacity Max	the maximum number of compute units that can be provided at peak times		Cloud service can be scaled during peak times. The capacity is the maximum number of compute units that can be provided at peak times.	D	M	C		X							X			X		X									X		S24
R	Resiliency	Maximum performance loss	MPL(%)=(Psteady - Pmin)/Psteady x 100 where Psteady and Pmin are the steady-state value of the considered performance index and its minimum value reached during the load burst, respectively.		measures the maximum performance loss. It is the peak of performance lost during the burst	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Response Time	Maximum Response Time(Tmax)	The maximum Response Time, denoted Tmax is the maximum Response Time claimed by the Cloud service provider.	time in milliseconds	Maximum Response Time is the maximum promised response time by the Cloud provider for the service.,	D	M	C		X							X	X		X	X	X										X	24,S30
R	Service Availability	Mean Availability	Am(T) = 1 /T * A(t)dt		The average uptime availability or mean availability can be defined as the proportion of mission time or time period that a system is available for use.	D	M	C					X				X			X	X	X									X		S52

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Performance for video streaming	Mean buffering duration (MBD)	bit rate of the video/ average bandwidth available for the viewer		The playback will pause when the amount of video data in the buffer is significantly low, causing a state of re-buffering. This is mainly due to the fact the bit rate of the video being greater than the average bandwidth available for the viewer	D	M	C					X				X	X							X						X		S64
PE	Capacity of a service	Mean number of jobs in RPDE (Erpde)	E[NRPDE] = ∑ number of jobs that are waiting in the RPDE queue + the job that is currently undergoing provisioning decision.		Mean number of jobs in the RPDE (E[NRPDE])is given by the sum of the number of jobs that are waiting in the RPDE queue and the job that is currently undergoing provisioning decision.	D	M	C					X				X			X		X										X	S28
R	PMs in active mode	Mean Number of PMs in active mode(Npm)	rm = Σ #Pm,active where #Pm denotes the number of tokens inside place P in marking m		The greater value for the mean number of PMs in active mode results in more power consumption of the VDC. Availability of the service considering all the customers of the VDC,	D	M	C					X				X			X	X										X		S62
PE	Number of requests	Mean Number of requests(q)	q = E|q|= ρ + λ2T2 (E[Y2]) / 2(1 − ρ)	number of requests	measures the mean number of requests in system (in waiting or in execution)	D	M	C			X						X			X				X								X	S47
PE	Performance of data centres	Mean queue length (MQL)	MQL = r + (r^c ρ / c!(1 − ρ)^2)P0 where C is the number of servers	number of the requests in the system	Mean queue length (MQL). The rate of the arrival and service requirements follows Pois- son processes.	D	M	C					X				X			X	X											X	S43
PE	Performance of data centres	Mean queue length (MQL)	MQL = Σ qi.MQLi i is teh number of operative servers		The MQL results are shown as a function of effects of the availability of the data centres in the cloud systems.	D	M	C					X				X			X	X											X	S43
PE	Response Time	Mean Response time (MRT)		Time in milliseconds. The lowest MRT indicates the fastest response time	The Mean response time represents the Round Trip Time (RTT) that took a ping of one CSP's virtual machine's address	B	A	C		X							X			X		X										X	S81
PE	Response Time	Mean response time(r)	mean Response time = Execution time +Waiting time, r = q/λ = x + λT2 (E[Y2])/ 2(1 − ρ)	Time in milliseconds	measures execution time can be calculated based on the total time for invoking the services less waiting time. Waiting time is the time taken for the request to be sent and receive a reply for the request.	D	M	C			X		X				X	X		X	X			X								X	47,S53
R	Mean Time between Failures	Mean Time between Failures(MTBF)	Reliability= MTBF = MTTF+MTTR MTBF is Mean Time between Failures, MTTF is Mean Time to Failure, MTTR is Mean Time to Repair	Time in minutes	measures the mean Time between Failures(MTBF)	D	M	C					X				X	X		X	X											X	S65

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Mean Time between Failures	Mean Time between Failures(MTBF)	MTBF= (Units*time)/number units failures		the ability of the cloud system to perform any required functions under the stated conditions for a specified period of time	D	A	C		X			X				X			X	X	X									X		S61
R	Mean Time between Failures	Mean Time between Failures(MTBF)		Time in minutes	measures the mean Time between Failures(MTBF)	D	M	C					X				X	X	X	X	X	X										X	83,S58
R	Mean Time between Failures	Mean Time between Failures(MTBF)	MTBF=Total time / Number of failures		MTBF applies to a service that is going to be repaired and returned to operate,	D	M	C	X	X			X				X	X		X	X	X	X								X		S09,S1C
P	Time Scaling	Mean time scale up		time in milliseconds	measures the mean time taken to expand or contract the service capacity. Elasticity of a Cloud service is inferred from the time it takes to scale up.	B	M	C		X							X			X		X									X		S24
M	Changeability	Mean Time to Change with respect to particular Cloud Service(MTTC)	Changeability(MTTC) = ∑(Time to analyze the change in workload + Time to modify the change in workload + Time to test the change in workload + time to distribute the change in workload)/(No. of change requests in workload)		Mean Time to Change with respect to particular Cloud Service(MTTC)	D	M	C					X				X	X			X											X	S65
R	Mean Time to Failure	Mean Time to Failure(MTTF)	MTTF =∫ tf dp (t) dt, where, ∫ fdp(t)dt = 1. Therefore, λ which represents "failure rate" λ=1/MTTF		measures the time period in which provider last in operation. The MTTF may be seen as a function of failure density	D	M	C		X			X				X	X	X	X	X	X									X	X	,S83,S6
R	Failure Recovery	Mean time to recover	MTTR = Avg (Time to a complete recovery from a service failure)	Time in minutes	measures the average time that a service takes to recover from a failure.	D	M	C			X		X			X	X	X		X	X	X		X							X	X	,S65,S6
R	Failure Recovery	Mean Time to Repair(MTTR)	MTTR	Time in minutes	mean time to repair(downtime)	D	A	C		X			X				X			X	X	X									X		S61
R	Mean time to switchover	Mean time to switchover		minutes	measures the time to switchover from a failure	B	M	C			X		X			X	X	X		X	X	X		X							X		S69
PE	Response time	Mean waiting time	w = r − x	Time in milliseconds	measures the average time for requests waiting in the queue	D	M	C			X						X			X				X								X	S47
PE	Memory capacity	Memory capacity		RAM size in GBs	measures the memory capacity.	B	M	C	X	X	X		X			X	X			X	X	X	X	X							X		S09,S10,S6
PE	VM Memory	Memory performance		MB/second	measures the property of storing data on a temporary basis, and it corresponds to the evaluation of the VM RAM	D	A	C		X							X			X		X										X	S81
PE	Memory utilization	Memory utilization	Memory utilization = current occupied memory resource /total memory resource * 100	ratio, The range is 0...100.	measures the percentage of current occupied memory resource in total memory resource	D	M	C					X				X			X	X	X										X	S35

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Mem usage container	memory_usage		ratio, The range is 0...100.	measures the percentage of Memory usage of container	D	A	C					X			X	X		X		X											X	S70
PE	VM Memory utilization	memUsed		size in GBs	measures the current memory usage of VM	B	A	C					X			X	X			X	X											X	S70
PE	VM Memory utilization	memUsedPercent		ratio, The range is 0...100.	measures the percentage of Memory usage of VM	D	A	C					X			X	X			X	X											X	S70
M	Service Modularity	Modularity	SM = No. of elements without external dependancy / total number of elements, denominator is the number of elements in a service, numerator is the number of elements which are not depend on external services.	If a service is entirely self-contained, SM will be 1	measures the service modularity	D	M	C					X				X	X			X										X		S53
PE	Server Capacity	Monitored process number		number of monitored process	measures the number of monitored process	B	M	C		X			X				X			X	X	X										X	S35
PE	Video Quality	Motion difference feature	Mn(i, j) = Fn(i, j) −Fn−1(i, j) Fn(i, j) is the luminance value of the pixel at the ith row and jth column of nth frame in time."		measures the difference between the pixel values (of the luminance plane) at the same location in space but at successive times or frames.	D	M	C					X				X	X							X							X	S66
PE	Capacity of a service	Net job rejection rate(preject)	Net job rejection rate(preject) = pblock + pdrop where pblock results from admission control, which results from rejection of jobs when RPDE buffer is full. Pdrop is rejection rate.		Net job rejection rate (preject)measured with two component of preject pblock and pdrop . 1. results from admission control and is denoted as pblock, which results from rejection of jobs when RPDE buffer is full. 2. rejection rate is denoted as pdrop.	D	M	C					X				X			X		X										X	S28
PE	Network rate	netBytesIn		Bytes/second	Bytes in per second	B	A	C					X			X	X			X	X											X	S70
PE	Network rate	netBytesOut		Bytes/second	Bytes out per second	B	A	C					X			X	X			X	X											X	S70
PE	Network rate	netPacketsIn		Packets/second	Packets in per second	B	A	C					X			X	X			X	X											X	S70
PE	Network rate	netPacketsOut		Packets/second	Packets out per second	B	A	C					X			X	X			X	X											X	S70
PE	Connections	New connections		The number of new connections	measures the new connections	B	M	C			X		X				X			X	X	X										X	S35
PE	Energy Efficiency	Node Energy efficiency	Ef = Useful work/Energy (J) = (Useful work/s)/(Energy (J)/s) = (Useful work/s) / Power (W)		measures the energy efficiency at node level for IaaS providers, is the ratio between the amount of work performed each second and the delivered power during that second	D	M	C					X				X			X		X										X	S41

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
M	Non-functional Commonality	Non-functional Commonality(NFC)	NFC = (∑number of requirements appling ith non-functional feature / total number of requirements analyzed in the domain)/m, where m is the total number of non-functional features in the domain. The non-functional features are identified from the SRSs in the same domain	The range is 0...1. The value 1 implies that all non-functional features are common in the same domain.	measures the average of commonality of each non-functional feature.	D	M	C		X					X		X	X			X										X		S44
S	Alarms monitored	Number of alarms monitored		Number of Fake alarms	Number of Fake alarms monitored by Corporate Security	B	M	C					X				X	X			X											X	S65
PE	CPU capacity VM	Number of CPU		Number of CPU cores(CPU_cores)	measures the computing using the number of CPU cores assigned to the monitored VM	B	M	C			X		X			X	X			X	X	X		X								X	51,S69
S	Ineffectual service	Number of ineffectual service responses		Number of ineffectual service responses	Number of ineffectual service responses to the issues identified by the Security as control weaknesses	B	M	C					X				X	X			X											X	S65
PE	Storage replication	Number of replicas		number of replicas	measures the minimum number of replicas guaranteed	B	M	C			X		X			X	X			X	X	X		X							X		S69
S	Safety hazards	Number of safety hazards		Number of safety hazards	Number of safety hazards proactively identified	B	M	C					X				X	X			X											X	S65
PE	Number of tenants	Number of tenants		Number of tenants	measures the number of tenants of a cloud service	B	M	C	X	X			X			X	X	X			X	X										X	S55
P	Operating systems support	Operating systems support		Operating systems	identify the different Operating systems Providers support, for example, Mac OS X, Windows and Open SUSE Linux.	B	M	Q	X	X	X	X	X		X		X	X	X	X	X	X	X	X							X		S10
R	Operational availability	Operational availability	A0 = Uptime /Operating cycle		operating cycle is the overall time of the investigated operation period whereas Uptime is the system's total functional time during the operating cycle.	D	M	C					X				X			X	X	X									X		S52
PE	Efficiency	Overall Efficiency(eff)	eff t-j = wav · avEff t-j +wrl · rlEff t-j +wrt · rtEff t-j w are the weights of availability, reliability, response time. Sum of weights is equal to 1		measures the overall efficiency regarding availability, reliability, response time	D	M	C	X	X							X	X			X		X								X		S45

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Overprovisioning	Overprovisioning	OVPR DT = 1 - 1/n where n is a number of the requests received and i the number of the request where the requested resources are beyond the available resources during a period of time DT.	Range 0..1. Value 1 implies that no request in over-provisioned. Value 0 means that all the request are over-provisioned	measures the over-provisioning of a cloud service	D	M	C		X							X			X	X											X	S02
P	Overprovisioning	Overprovisioning resource	∑ over provisioning resource		measures provisioning resource is over the needs, the SLA and QoS are not optima	D	M	C					X				X			X	X	X										X	S35
R	Packet Loss	Packet Loss Distribution (PLD)		PLR/sec	measured by PLR per time	D	M	C		X			X				X			X	X	X	X					X					S23
R	Packet Loss	Packet Loss Rate (PLR)	P prob = 1/N Σ n k where Pprob is the probability of packet loss, N is total number of packets, nk is number of lost packets by the k-th period	The range is 0...1	Packet loss probability – defined as the percent number of lost packets to the total number of packets, ratio between number of lost and sent packets	D	M	C		X			X				X			X	X	X	X					X					S23
R	Packet loss	Packet losses		number of packets	measures the amount of packets lost in an interval. A discontinuity in the sequence number of RTP packets is perceived as a loss. Packet loss causes degradation in voice and video quality.	B	M	C					X			X				X	X										X		S16
PE	Packets in	Packets in		number of packets	measures the number of packets in	B	M	C					X				X			X	X	X										X	S35
PE	Packets out	Packets out		number of packets	measures the number of packets out	B	M	C					X				X			X	X	X										X	S35
PE	Web Services Performance	Page download rate(Rate)	Rate(p)=Bytes(p) / Page(t), where Bytes(p) is the bytes of the request page, Page(t) is the download time of request page.	rate	measures the page download rate,	D	A	C					X				X	X	X	X	X				X							X	S08
PE	Participant joining time	Participant joining time		Time in milliseconds	measures the time required to add a participant to a running conference	B	M	C					X			X	X	X			X											X	S04
R	Resiliency	Peak Overshoot or Undershoot (PO)	PO(%) = (|Mpeak(ac) −Mss(ac) | / Mss(ac))× 100% where Mpeak(ac) be the maximum deviation of the measure from its steady state value after application of change.	percentage	measures the Peak Overshoot or Undershoot (PO)	D	M	C					X				X			X		X										X	S28
FS	Peak Signal	Peak Signal to Noise Ratio	PSNR	dB	Peak Signal to Noise Ratio in Y component	D	M	C		X			X				X			X	X	X	X					X					S23

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Resiliency	Peak Time		time in milliseconds	Peak Time (Tpeak) is the time taken to reach maximum deviation Mpeak(ac) of the measure, from its steady state value (Mss (ac)) attained after the change is applied	B	M	C					X				X			X		X										X	S28
S	Dangerous data	Percentage of dangerous data resources		percentage	Percentage of dangerous data resources residing on systems	D	M	C					X				X	X			X											X	S65
PE	Capacity of a service	Percentage of QoS violations	Percentage of QoS violations: percentage of requests whose response times were longer than the amount of time expected by their end-users;	range 0...100	measures the percentage of requests whose response times were longer than the amount of time expected by their end-users;	D	M	C					X				X			X	X				X							X	S18
PE	Response Time	Percentile of response time	FT(t)|t=TD= ∫fT(t)dt ≥ γ (0 ≤ γ≤ 1) Let T be a random variable representing the response time, fT(t) and FT(t) be its probability and cumulative distributions pdf and CDF, respectively. TD be the desired target response time that a customer requests and agrees with its service provider based on a fee paid by the customer. which is called percentile response time. This means that γ×100% of the time a service request job will be executed in less than TD. , F(t) =1-e^rT =− ≥σ the probability of the response time being less than T B should be no less than σ.	Time in milliseconds	measures the Percentile of response time	D	M	C					X				X	X	X	X	X	X			X							X	78,S77

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Resiliency	Percentile Time	tpercentile=|Mss(bc)−Mss(ac)|. (100 − γ)	percentage	γ-Percentile Time (tpercentile). This is useful when the measure of interest never goes above or below the steady state value (Mss(ac)), after the change is applied. It is defined as the elapsed time from the time when the change is applied to the system until the output measure reaches is always greater than δ as used in settling time definition	D	M	C					X				X			X		X										X	S28
PE	Performance of service	performance	performance=< Time Reliability Cost >		measures the performance	D	M	C					X				X			X	X											X	S48
PE	Performance of service	performance		number of accomplished requests/s	measures the number of accomplished requests per unit time	D	M	C					X				X			X	X				X							X	S79
PE	Degradation due to Migration	Performance Degradation due to Migrations (PDM)			measures the performance degradation due to migrations (PDM). is bounded by host capacity and VMs are not provided with the required performance level	B	M	C					X				X	X				X			X							X	S25
PE	Performance utility	Performance utility function	UP (X)= {uXmin + (1 + uXmin) .|RXP-X / RXP-IXP| where IXP ≤ X ≤ RXP, otherwise: 0, uX min is the minimum utility that the consumer and the provider receive for reaching a deal at its reserve performance. Qp be the performance quality at which a consensus is reached by both parties. Qp is substituted to X which are generalized utility functions for cloud QoS so that the performance utility is UQp (Qp) for reaching a consensus at Qp where IX=IQp, and RX=RQp	ratio, The range is 0...100. consumers prefer the highest guarantee for the service performance, and providers want to sell their services with the lowest guarantee for the service performance with a given price	represents the level of satisfaction about the provisioned resource constraints, throughput, or response time. The consumer and the provider who provides the VM service can negotiate level of the guarantee for the performance resource constraints.	D	M	C		X							X	X		X	X	X									X		S67

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Performance utility	Performance utility function	for each component cj ∈ C, we define a function perfj : (λ, ρj) ?→ N in terms of response time, where λ corresponds to the workload and ρj to the total amount of CPU allocated to component cj. One application is composed of a set of components expressed by the vector C.		revenues of a given application configuration k running under a certain infrastructure configuration N. Indeed, the higher the response time the lower is the revenue.	D	M	C					X				X			X	X										X		S06
PE	Performance of Virtual Machine	performance VM	δ(x) = ηe(1 − ptime−out) defined as δ(x) with x representing the number of available VMs. δ(x) gives the number of completed tasks per unit	number of completed tasks/unit	reflects the impact of the number of available VMs on the service performance.	D	M	C					X				X			X	X				X							X	S79
PE	Performance of the infrastructure	Performance-Cost Normalization (PCN)	Performance-Cost Normalization (PCN) = P/C, where C is execution cost, P is the parameter to measure the performance of the reference application(e.g. response time).		measures a running application or a software program on a cloud service by determining the cost incurred during the process and normalizing the output of the application or program by the cost. PCN helps you decide on the right provider that makes sound business sense.	D	M	C	X	X							X			X		X	X								X		S71
P	Server vertical scalability	Permissible capacity fluctuations to workloads		number of CPUs, RAM size in GBs	measures permissible capacity fluctuations to workloads. fluctuation number of CPUs and RAM size in GBs.	B	M	C			X		X			X	X			X	X	X		X							X		S69
P	Server horizontal scalability	Permissible changes to increased workloads		number of virtual servers in resource pool	measures permissible changes to increased workloads	B	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Persistence	Persistence			It is defined as the number of time periods required for a given proportion of the total uncertainties in a given service to collect	D	M	C					X				X	X		X	X											X	S65

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
FS	Pertinence of service	Pertinence value	where ç is the service unit (processing, network, storage, etc.), nç is the number of expected values, çu-i is the service unit value expected by user u, and is the service unit value sent back to the user u		measures the pertinence value	D	M	C		X							X	X			X	X									X		S30
P	Platform support	Platform support		Platforms	identify the provider different types of platforms.	B	M	Q	X	X	X	X	X		X		X		X		X	X	X	X							X		S10
PE	Energy Consumption	PM Energy consumption	Ej = Ej min + ut CPUj . (Ej max - Ej min) Ej min and Ej max represents minimum and maximum energy consumption of a particular PM j, ut signifies the utilization of the CPU of PM Physical Machine	Energy consumption	measures the energy consumption of a physical machine (PM)	D	M	C		X							X			X		X									X		S38
P	Portability ratio	Portability of service	Portability= (number of compatible platforms)/(total number of platforms)	ratio, The range is 0...100.	measures the degree to which the service or CS is portable to other platforms	D	M	C					X				X	X			X											X	S65
PE	Power Consumption	Power Demand		watts	measures the total watt usage by the cloud system, Power consumption of a cloud computing system (MWatt)	B	M,A	C		X			X				X		X	X	X	X						X			X		S61,S36
PE	Performance of matching algorithms	Preparing Time(Tp)	TP = tf − ts where tf time first and tl time last matched subscription	Time in milliseconds. Low preparing time means low matching delay to forward events to subscribers.	Preparing time is the common matching delay for all matched subscriptions.	D	M	C					X				X	X							X							X	S56
S	Privacy	Privacy	Privacy= (Ac_state,Ac_trans)		describes privacy protection in cloud, including access control state and its state transition	D	M	C					X				X			X	X											X	S48
S	APT Accesed	Probability an APT accessed			ATP can access high value data. Data compromised or erased from the CCS	D	M	C					X				X			X	X	X										X	S29
S	APT Detected	Probability the APT detected	Pr DIN = (Pr (A | nodetection) - Pr(A))/Pr(A) (Pr (A | nodetection) is the probability that the APT can infiltrate the gold data without any detection (i.e., Pr DIN=0 for i = 1; ... ;n.		APT is detected by cloud tenant or CCS security monitoring systems. Intrusion detection.	D	M	C					X				X			X	X	X										X	S29

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Processing time	ProcessingTime		Average request processing time	measures the processing time	B	A	C					X				X	X			X											X	S70
PE	Processing time	ProcessingTime	average response time	time in milliseconds	average response time	D	M	C	X	X			X				X			X	X	X		X							X		S26
PE	Productivity	Productivity	P, (Eq. (4))		Cloud service performance per unit cost, (TFlops/$, WIPS/$, etc.)	D	M	C	X	X			X				X	X		X	X	X						X					S36
PE	Turnaround Time	promised turnaround time		time in milliseconds	The promised turnaround time is the expected time by a resource provider between the submission of a job and the delivery of the completed job. It is promised by the resource provider to the user in the SLA.	B	M	C		X							X	X	X	X	X	X									X		S50
P	Provisioning time	Provisioning time	It is defined as the time taken by a CSP to allocate a new VM once the customer requested it.	time in milliseconds	measures the sum of the resource allocation time for the VM request, and the time to power on the VM to get it ready for use. The provisioning time has a direct impact on the scalability of one cloud based application. scaling latency.	D	M	C		X							X			X		X										X	S81
PE	Video streaming	Quality shifting frequency(QSF)		number of changes (upshifting or downshifting)	This indicator measures how frequently the resolution shifting occurs in the video player	D	M	C					X				X	X							X						X		S64
R	VM Availability	Ratio of available VMs to accepted VM requests(Rava)	rm = (Σ #Pm,vmsi)/(#Pm,vtp + #Pm,vtm + Σ (#Pm,vmsi + #Pm,vtpi + #Pm, vtmi)) where #Pm denotes the number of tokens inside place P in marking m. In the states in which the denominator of the fraction is zero, the reward rate is 1,		the mean ratio of the number of available VMs to the total number of accepted VM requests in steady state. cloud provider wishes to minimize the power consumption of the VDC.	D	M	C					X				X			X	X										X		S62

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Continuity	Readiness of emergency preparedness for key businesses (RoEPKB)	k =v / 2 (6) k = RoEPKB, 0≤k≤1; v = 2 when service provider has established comprehensive mechanism on emergency preparedness for key businesses and can fulfill the specification well; v = 1 when service provider has established mechanism on emergency preparedness for key businesses, but cannot fulfill the specification very well; v = 0 when there is no mechanism on emergency preparedness for key businesses		The closer the value of "k" is to 1, the better the RoEPKB.	D	M	Q	X						X			X	X	X					X							X	S84
PE	Latency	readLatency		Time in milliseconds	Keyspace read latency	B	A	C					X				X	X			X											X	S70
PE	Video streaming	Re-buffering frequency (RBF)	Re-buffering frequency (RBF) = Avg (buffering duration)		measures the average buffering duration which takes into account the initial buffering time as well as the re- buffering events	D	M	C					X				X	X							X						X		S64
PE	Response Time	Recommended Service Response Time(RSRT)			measures the Recommended Service Response Time(RSRT)	D	A	C		X							X	X			X	X										X	S17
R	Resiliency	Recovery rate	Rc	percentage	measures the recovery rate or the capability to recover from failure or disaster (%)	D	M	C					X				X				X	X						X					S36
R	Recovery point objective (RPO)	Recovery point objective (RPO)		size in MBs	amount of data to be lost as a result of a fault or disaster,	B	M	C					X				X			X	X	X									X		S52
R	Resiliency	Recovery time	Trec = tr - tf where the time Trec elapsed between the burst end (tf) and the instant (tr) at which the performance indices returns to a value greater than the α-100% of its steady-state value.		measures the recovery time. Is the time elapsed between the burst end (tf) and the instant (tr).	D	M	C		X	X						X			X	X	X		X								X	S14
R	Recovery time objective (RTO)	Recovery time objective (RTO)		time	minimum downtime for recovering from faults	B	M	C					X				X			X	X	X									X		S52

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase: Req.	Adq.	Dev.	Int.	Ope.	Ret.	Artifact: Spe.	Arc.	Ser.	Service Type: SaaS	PaaS	IaaS	Stakeholder viewpoint: CSP	CSC	CSB	CSD	USE	Validation procedure: TheoreticalValidation: A.A.	A.T.M	Empirical Validation: C.E.	C.S.	SU.	No Validation: N.V.	P.C.	Ref. Paper
PE	Request Rejection	Rejection probability(RJ)	RJ=1-(1-RJC) * (1-RJP) RJC is the request rejection probability at the CMU handling phase, and is the expected number of jobs that one request generates. RJP job rejection probability		measures the probability that a newly arrived request is rejected at the CMU handling phase	D	A	C					X				X			X					X						X		S76
P	Reactivity	Release reactivity	Re = 1 - Σ Min(1, Tr-Tp/(Tp - Ta)) / K where Ta: allocation time, Tp: procesing time, Tr: release time.	Range 0..1	reflects the reactivity of the cloud against the workload	D	M	C		X							X			X	X				X							X	S02
R	Service Reliability	Reliability	R(t)= -t/e^MTBF		measures the probability that a service operating under specified conditions shall perform satisfactorily for a given period of time	D	M	C	X	X			X				X			X	X	X									X		S09,S10
R	Failure rate	Failure rate	the number(or time) of failures / the mean number (or time) of failures promised by a provider in a given time interval	Positive.	Reliability represents the ability of a service to perform its required functions under stated conditions for a specified time interval	D	M	C		X			X				X	X	X	X	X	X						X					S49
R	Success Operation	Success rate	Reliability of resource Rk(RE) = Ck / Ak, Ak number of works/jobs accepted by Rk resource Ck number of works/jobs completed successfully by Rk source in T time limit		the ability of a system or component to perform its required functions under stated conditions for a specified period of time—IEEE, is the ability of a system or a component required for operation under steady-state conditions for a specific time period. Measure of success of a work accepted by the cloud supplier. Success rate	D	M	C		X			X				X	X	X	X	X	X										X	S50,S27
R	Service Reliability	Reliability	Reliability = Pr[V]×P[TF] =(1 − nf/n) P[TF] Pr[V] is the probability of violation, nf is the number of users who have experienced a failure, n is the number of users in a concerned time period, and P[TF] is the promised mean time to failure.		Reflects how well a service operates without failure during a given time and condition. Also the ability to ensure a continuous process of the program without loss and is characterized by the number of failures. Mean time to failure promised by the provider which is the length of time a device or an application is expected to last in operation.	D	M	C	X	X			X				X			X	X	X									X	X	7,S26,S

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Service Reliability	Reliability	Rel(Ri)={ 1, RRi ≤ VRi; (1- exp(- VRi))/(1-exp (-RRi)), otherwise where RRi is Reliability of request Ri and RRi ϵ (0,1), and VRi is reliability of virtual machine who responses the request Ri, and it is also the lowest one of all virtual machines that satisfy the request Ri.		Reliability is related to the number of requests and virtual machines	D	M	C					X				X			X	X	X									X		S75
R	Service Reliability	Reliability	Reliability= W.CFT +W.CFR + W.SA where w is the weight for each metric, which the sum is 1 Coverage of Fault Tolerance (CFT) + Coverage of Failure Recovery (CFR) + Accuracy (SA)	The range is 0...1 Higher value indicates that the SaaS is more reliable.	measures the reliable of the service	D	M	C		X			X				X	X			X										X		S44,S53
R	Service Reliability	Reliability	ρ=1-n/ni , p = 1 - n/ns n and n denote the s number of failed and total operations that occurred in a time interval	The range is 0...1 The closer the value of is to 1, the higher the reliability	REL Evaluation: REL means the assurance that a storage cloud is free from hardware failures, software faults, network outages and other defects that could make them break down.	D	M	C					X				X	X	X	X					X			X	X			X	S82,S83
R	Service Reliability	Reliability	Re(t) = e−λt		Re, represents the capacity of a service to achieve its functions under certain conditions and in a given period of time	D	M	C		X							X	X			X	X									X		S30
R	Reliability Efficiency	Reliability Efficiency	rlEfft = rlSaaS t-j / (Σ k∈Kt rlCloudt / |Kt-j|) K t-j is the set of cloud providers that are servicing SaaS provider j at time t		defines how well a SaaS provider can provide reliability to a user based on the reliability it receives from the cloud provider	D	M	C	X	X							X	X			X		X								X		S45
R	Reliability Importance	Reliability Importance	RIi(t) = ∂Rs(t) /∂Ri(t) where Rs(t): is the system reliability at a certain time, t; Ri(t): is the component reliability at a certain time, t.		identify the relative importance of each component according to identifying the importance of each system component.	D	M	C					X				X			X	X	X									X		S52
R	Reliability of Links	Reliability of Links	R DL KL(t) Where t D kl (X) is duration in which DL kl is used to transmit data		reliability of system without considering failure of servers can be computed by multiplying reliability of the links	D	M	C					X				X			X	X	X									X		S22

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Reliability of processor	Reliability of processor	R Pj (X) = e^-λj ∑ Eij Xij Reliability of processing node Pj in time interval [0,t]. As the total time for executing the tasks assigned to processor Pj by assignment X		As the total time for executing the tasks assigned to processor Pj by assignment X	D	M	C					X				X			X	X	X									X		S22
R	Reliability state	Reliability state	Reliability: Trans--> Real		measures the reliability of the state transition	D	M	C					X				X			X	X											X	S48
R	Service Reliability	reliability utility function	time slot utility UPT(X) = uX,min + (1 + uX,min)* |RXp -X / RXp-IXp| ; IXp ≤ X ≤ RXp, otherwise 0 Qr be the reliability performance quality at which a consensus is reached by both parties. Qr is substituted to X which are generalized utility functions for cloud QoS so that the performance utility is UQa(Qa) for reaching a consensus at Qr where IX=IQr,and RX=RQr and be the most preferred (initial) service quality and the least preferred (reserve) service quality for a consumer agent (a provider agent),	ratio, The range is 0...100. consumers prefer the highest guarantee for service reliability, and providers want to sell their services with the lowest guarantee for service reliability with a given price	Reliability is the likelihood of successfully using a service.	D	M	C		X							X	X		X	X	X									X		S67
R	Continuity	Repair rate of accidents (RRoA)	r = a / b; r = RRoA, 0≤r≤1; a = the number of accidents repaired within required time in service agreement; b = the total number of accidents within a service period	The range is 0...1 The closer the value to 1, the better the RRoA.	The closer the value of "r" is to 1, the better the RRoA.	D	M	C					X				X	X	X	X		X										X	S84
PE	Server Capacity	Replication		number of replicas	measures the number of replicas	B	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Performance of service	requestCount		number of requests	measures the number of requests served since last collection	B	A	C					X				X	X			X											X	S70

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Resource Scaling	resource (r)	r=ϕ× AVEd + (1- ϕ) × AVEs, ϕ∈ [0,1] where Average Overprovisioning resource (AVEs), Average Underprovisioning resource (AVEd)		measures the resource Scaling	D	M	C					X				X			X	X	X										X	S35
PE	resource allocation	resource allocation	total amount of allocated resources (RAM, cpu)		measures the total amount of allocated resources to accommodate all participants.	D	M	C					X			X	X	X			X											X	S04
PE	Capacity of a service	Resource in use	wtb=Number of resources in use at time t		wt represents the number of resources in use to handle requests at time t; note that wt< mt indicates that mt− wt active resources are not being used by the service provider, mt represents the resource pool capacity	D	M	C					X				X			X	X				X							X	S18
PE	Capacity of a service	Resource pool capacity	mt = Service provider's resource pool capacity at time t		The service provider allocates up to m resources from the cloud provider in order to host its service and handle end-users' requests, and its resource pool capacity at time t ∈ T is denoted by mt for each t ∈ T; recall that, with auto-scaling, mt may be different from mt' for t≠ t'	D	M	C					X				X			X	X				X							X	S18
PE	Resource Utilization	Resource utilization	UoR = RU = (amount of allocated resources) /(amount of pre - defined resources) where the denominator is the amount of pre-defined resources for each type of resources and it is defined in the SLA.	The range is 0...1 Higher value indicates that the SaaS has higher resource shares	measures a ratio of an amount of allocated or assigned resources for the pre-defined resources. The resources include a network bandwidth and other computing powers such as storage capacity, CPU.	D	M	C		X							X	X			X										X		S44,S53
PE	Resource Utilization	Resource utilization	Resource utilization(ut) =number wt of resource hours effectively used to handle requests/ number mt of resource hours allocated from the cloud (i.e., idle resources are not considered)		measures the resource utilization at time t, is the ratio between the number wt of resource hours effectively used to handle requests and the number mt of resource hours allocated from the cloud (i.e., idle resources are not considered)	D	M	C					X				X			X	X				X							X	S18

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Resource Utilization	Resource utilization	U= resource allocated VM/resource size	A high value of resource utilization is a desirable behavior expected from a cloud-based service delivery system.	Resource utilization refers to the fraction of resource size that is allocated to serve the user requests. These resources are the virtual machines hosted on top of the physical machines and the utilization level of the virtual resources of each virtual machine is called as resource utilization.	D	M	C					X				X			X	X				X						X		S42
PE	Response Time	Response mean time (synchronous operation)	Mean time of response: avg in milliseconds	time in milliseconds A low value of a response time is a desirable expected from a cloud-based service	measures the response time in milliseconds for a request to be serviced	D	M	C	X	X	X		X			X	X			X	X	X	X	X							X		S69,S71
PE	Response Time	Response mean time (synchronous operation)	response time= time request service is available - time of request for service	time in milliseconds A low value of a response time is a desirable expected from a cloud-based service	measures the difference between time of request for service and time when service is available, can be measured by benchmark	D	M	C	X	X	X		X			X	X			X	X	X	X	X							X		S09,S10
PE	Performance of the infrastructure	Response mean time	Mean time of response: avg(time for a request to be serviced)	time in milliseconds	measures the response time in milliseconds for a request to be serviced	D	M	C	X	X	X		X			X	X			X	X	X	X	X							X		S69,S71
PE	Response time	Response standard deviation time (synchronous operation)	Standard deviation time of response: avg in milliseconds		measures the Response standard deviation time (synchronous operation)	D	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Response Time	Response Time	response time = time user receives a response from a cloud - the time a user sends a request to the cloud Time between sending a request and getting a response	time in milliseconds The tendency of the response time measurement is low, meaning that the lower the value the better the service performance is.	measures the time it takes when a user sends a request to the cloud and receives a response from a cloud solution. This includes the processing time at the SaaS and IaaS levels and network transmission time.	D	M	C	X	X			X				X	X	X	X	X	X			X			X			X	X	S42,S49

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Response Time	Response Time		time in milliseconds A low value of a response time is a desirable expected from a cloud-based service	Is the time for a service request to be satisfied. That is, this is the time it takes for a service request to be executed on the service provider's multiple resource sites	D	M	C					X				X	X	X	X	X	X										X	S78
PE	Response Time	Response time		time in milliseconds	measures response time of the application when executing all the queries.	B	M	C			X						X	X						X				X					S33
PE	Response Time	Response Time	time elapsed from a user's submitting a service request until the final output is received RT = Tf −Ts		measures the Response Time	D	M	C	X	X			X				X			X	X	X		X							X		S26
PE	Response Time	Response Time	Time since it provides better approximation of entire distribution than the average response time	time in milliseconds	Map of response time and workload arrival rate to compare with the target set-point.	D	M	C					X				X	X			X	X										X	S31
PE	Response time	Response time	time it takes to receive the first byte of any file being requested.		the time it takes for a request made by a client or workload generator to be received by the client. Ping response time between client and server	D	A	C		X			X				X			X	X	X									X		S61
PE	Response Time	Response Time	The value of maximum time response expected for processing the input workload		measures the Response Time	D	M	C					X				X		X		X										X		S68
PE	Response Time	Response Time	RT=delivery time= delivery of done work - request time		Response Time is the exact time between requested time of work and the delivery time of work to the user	D	M	C					X				X	X	X	X		X											S27
PE	Response Time Efficiency	Response time efficiency	rtEff t-j = max k∈K t-j rtCloud t-k / rtSaaS t-j K t-j is the set of cloud providers that are servicing SaaS provider j at time t		defines how well a SaaS provider can provide response time to a user based on the response time it receives from the cloud providers	D	M	C	X	X							X	X			X		X								X		S45
PE	Response Time	Response Time Failure	100(n′/n), where n′ is the number of occasions when the service provider was not able to fulfill their promise.		Response Time Failure is given by the percentage of occasions when the response time was higher than the promised maximum response time.	D	M	C		X							X			X		X									X		S24

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Response Time	Response Time Failure	σrep = n\`/n * 100 where n is the total number of service accesses, n\` the number of times where the Response Time exceeds the expected maximum value (Trepi> Tmax,repi)	The range is 0…100	Response Time failure is expressed as a percentage of cases exceeds the maximum Response Time Tmax,repi promised by the service provider.	D	M	C		X							X	X			X	X										X	S30
PE	Response Time	Response Time(Trep)	Treq-rep = 2 × Treq-net (n) + Treq-net	time in milliseconds A low value of a response time is a desirable expected from a cloud-based service	The Response Time is the sum of the sending time (of both the request and the response) and the processing time of the request.	D	M	C		X							X	X			X	X										X	S30
M	Reusability	Reusability	Reusability = W.UoS+ W.AoS + W.CoV + W.CF The weights for each metric which the sum is 1. Understandability of Service (UoS), Awarability of Service(AoS), Coverage of Variability(CoV), Commonality feature(CF)	The range is 0...1 and the higher value indicates the SaaS is more reusable	Reusability indicates that the web services is more reusable	D	M	C					X				X	X			X										X		S53
M	Reusability	Reusability	Reusability = w.(FC) + w.(NFC) + w.(CV) where w is the weight for each metric, which the sum is 1. Functional Commonality(FC)+ Non-functional Commonality(NFC) + Coverage of Variability(CV)	The range is 0...1 and the higher value indicates the SaaS is more reusable	measures the degree of reusable service	D	M	C		X							X	X			X										X		S44

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Operational availability	Robustness of Service	ROS=(available time for invoking SaaS) / (total time for operating SaaS) where the numerator is the specific period of time to be able to invoke the SaaS and the denominator is a total period of time to start operating the SaaS The numerator can be calculated as 'total operating time of SaaS – Service Failure Time'. The Service Failure Time is a total time when the service cannot be invoked because of any service failures.	The range is 0...1 The higher value indicates that the SaaS has higher availability.	measures a ratio of the total operating time to the time which the SaaS is available to be invoked	D	M	C		X							X	X			X										X		S44
R	Robustness	Robustness	Rob = Σny / nH the average number of services on hosts:	The more the number of service affected is small, the more the Cloud provider can insure to these users a high level of robustness	the probability of a service to be affected by a failure of a component of the Cloud. in the provider point of view the number of services affected by an host failure.	D	M	C		X							X	X			X	X										X	S30
PE	Latency	Round-trip reaction delay (RTRD)	Round-trip reaction delay = transmission delay + server-side processing delay + client-side processing delay		latency is measured in terms of round-trip reaction delay (RTRD), which is defined as the elapsed time between the event of keystroke from a user and the display of the first corresponding video frame affected by this keystroke.	D	M	C					X				X	X						X								X	S74
PE	CPU Performance	RunSpeed Value	RunSpeed Value= CPU job / CPU resource		measuress the RunSpeed Value	D	M	C					X				X	X	X	X		X											S27
PE	Bytes received	rx_bytes		bytes	measures the bytes received (rx_bytes)	B	A	C					X			X	X		X		X											X	S70
PE	Packets received	rx_packets		Packets received	measures the packets received	B	A	C					X			X	X		X		X											X	S70
PE	Energy Efficiency	SaaS Energy efficiency	Ef Servicek = Σ j∈Service EfVMj / mk	Energy consumption	measures the energy efficiency at service level for SaaS providers. Energy efficiency of a given service would be to evaluate the mean energy efficiency of all the VMs (m) hosting it.	D	M	C					X				X	X				X										X	S41

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure								Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Energy Cost	Saving Cost	The saving cost on energy can benefit users with discount price, calculated as $\Delta C = cu \cdot \Delta E$ and $\Delta E = Einit\ E{-}opt$, where cu is the unit price of energy charge, EΔ is the saving energy, Einit is the energy consumed by completing the application tasks with the initial resource allocation, and Eopt is the energy consumption consumed by completing the application tasks after VM scheduling		measures the saving Cost on energy	D	M	C					X				X	X							X							X	S19
P	Scalability	Users Scalability		the number of users	the number of users it can support from a baseline value.	B	M	C					X				X			X	X	X										X	S58
P	Scalability	Scalability rate	maximum available increase capacity / time interval		Scalability represents the capability of increasing the computing capacity of service provider's computer system and system's ability to process more users' requests, operations or transactions in a given time interval.	D	M	C		X			X				X	X	X	X	X	X						X					S49
P	Scalability	Resource Scaling	S (Eq. (5))		The ability to scale up resources for gain in system performance This measure is inversely proportional to the service costs and directly proportional to productivity. This metric evaluates the economy of scale by a pair of productivity ratio.	D	M	C					X			X	X	X	X	X	X	X	X					X					S36
P	Scalability	service Scaling	S = T srep/n req where Response Time of a service (T srep) in terms of number of requests nreq the service receives		measures the Scalability of the service	D	M	C		X							X	X			X	X									X		S30

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Time Scaling	Scalability Time	Tm = Tj +Tu +To where Tj : time just needed Tu: time underprovisioning To: time overprovisioning	time in milliseconds	total measuring time including all the periods in the states above: Tu is the under-provisioning time that the cloud platform needs to switch from an under-provisioning state to a corresponding balanced state and To denotes the over-provisioning time in which cloud platform needs to switch from an over-provisioning state to a balanced one in those periods of time.	D	M	C					X				X			X	X	X									X		S26
S	Security level	Security	secure =<Integrity, Privacy, Availabilty, Authenticity>		measures security level base on Integrity, Privacy, Availabilty, Authenticity	D	M	C					X				X			X	X											X	S48
S	Security mechanism	Security mechanism		List of security mechanism available	The security in Cloud computing means a set of technologies and user policies designed to follow the regulatory compliance rules and to protect all the resources like information, applications and hardware infrastructure connected with cloud computing use.	B	M	Q	X	X	X	X	X		X		X	X	X	X	X	X	X									X	10,S58
S	Security breach	Security (SECY-CQ)	θ= 1- FT(t) where θ represents security and FT(t) denotes a cumulative distribution function of a random variable T indicating the time until the first security breach occurs	time	SECY Evaluation: SECY denotes the assurance that users' data stored in a storage cloud is under protection and free from data leaks., the assurance that Cloud services are free from viruses, intrusions, spyware, attacks, and other security vulnerabilities that could put them at risk.	D	M	C					X				X	X	X	X					X			X	X			X	S82,S83
S	Security cost	Security cost	Security cost = % of total company revenue	percentage	measures the security cost as a percentage of total company revenue	D	M	C					X				X	X			X											X	S65
S	Security index	Security index value (Iv)	Index = Σ(SV Fi) W(Fi Cd) /n * 100 - n represents the total number of factors considered in the security index calculation - SV Fi = Σ SF = SF + SF+... + SF where m represents the total number of subfactors for a given factor	The value of subfactor can be bounded as: 0 ≤ SF ≤1		D	M	C		X							X	X	X	X					X							X	S60

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
S	Security Management	Security Management	This is a rating of the extent to which a service can satisfy user security requirements in terms of access control, privacy, data, infrastructure, etc		measures the rating of the extent to which a service can satisfy user security requirements	D	M	Q		X			X				X	X			X	X										X	S21
S	Security level	security metric	Na(x, n)(x ≤ n) n is the total number of VMs and x is the number of available VMs.		reflects the impact of security factors on the number of available VMs.	D	M	C					X				X			X	X				X							X	S79
S	Security level	Security(Sec)	Sec(Ri)={ 1, RSi ≤ VSi; 1- (RSi -VSi)/highestlevel - Vsi, otherwise where RSi is security level of request Ri, and VSi is security level of virtual machine who responses the request Ri.	Highest level is the highest security level in the system.	indicates the security level to represent the security of both requests and virtual machines. Set four security levels: {poor, low, medium, high}.	D	M	C					X				X			X	X	X									X		S75
PE	Energy Consumption	Server Communication Energy consumption	∑ESCom= ∑ ∑ Dpq/bpq·wpq where Dpq is the amount of data transferred from server Sp to Sq, bpq is the communication bandwidth between Sp and Sq, wpq represents the energy consumption rate for the communication between Sp and Sq.	Energy consumption	measures the energy consumption for communication ∑ ESCom among all servers	D	M	C					X				X	X							X							X	S19
PE	Energy Consumption	Server Energy consumption	ES = EBase + ∑ERun + ∑EIdle + ∑EVCom where EBase is the basic energy consumption of S, ∑ERun is the energy consumption of all running VMs in the server, ∑EIdle is the energy consumption of all idle VMs in S, ∑EVCom is the energy consumed by communication among all VMs in S	Energy consumption	measures the Server Energy consumption	D	M	C					X				X	X							X							X	S19

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
R	Server reliability	Server reliability	reliability of each server can be calculated by multiplying its processors and other components reliability		for modeling server reliability, memory, hard disk, RAID controller and processor reliability of each server should be determined	D	M	C					X				X			X	X	X									X		S22
P	Scalability	service Scaling	$s = \lambda \times AVT_u + (1-\lambda) AVT_d$, $\lambda \in [0,1]$ where average delay time at scaling down (AVT_d), average delay time at scaling up (AVT_u)	time in milliseconds	measures the service Scaling	D	M	C					X				X			X	X	X										X	S35
PE	Execution time	Service local Time	$Sl = E[r_{ux}] / Thr[T_{serv}]$ where E is expected number of running VMs and Thr is the expected throughput of transition Tserv	Time in milliseconds	Sl is the time needed to locally execute a VM. It measures the ratio between the expected number of running VMs and the expected throughput of transition Tserv	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Execution time	Service Remote Time	$Sr = l/\mu * q_f$ where qf is federated quality level. It is used when VM multiplexing is not allowed.	Time in milliseconds	Sr is the time needed to execute a VM into the federated clouds.	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Execution time	Service Time	$S = Sl\,(E[rl\text{-}x]/(E[rl\text{-}x]+E[rr\text{-}x])) + Sr\,(E[rr\text{-}x]/(E[rl\text{-}x]+E[rr\text{-}x]))$ Sl and Sr have to be weighted summed with respect to the conditional probability of being locally or remotely accepted	Time in milliseconds	measures the expected time S needed to execute a VM and it can be obtained locally or remotely in federated cloud	D	M	C		X	X						X			X	X	X		X								X	S14
PE	service time of the task	service time of the task	$f_e(t) = \mu_i \cdot e^{-\mu_i t}$		measures the service time of the task	D	M	C					X				X			X	X				X							X	S79
R	Resiliency	Settling Time		time in milliseconds	Settling Time (tset). This is defined as the elapsed time from the time when the change is applied to the system until the measure of interest reaches and stays within ±δ% of $\|M_{ss}(bc)-M_{ss}(ac)\|$	B	M	C					X				X			X		X										X	S28
PE	Capacity of a service	Shortage capacity	$Shortage(i) = Cap\_min(i) - Cap\_prov(i)$ when $Cap\_min(i) > Cap\_prov(i)$ and zero otherwise		measures the shortage capacity for interval i	D	A	C		X			X				X			X	X	X									X		S61
PE	Skew	Skew		time in milliseconds	Clock skew between sender and receiver from RTP timestamp	B	M	C		X			X				X			X	X	X	X					X					S23

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	SLA violation	SLA Violation Time per Active Host (SLATAH)	SLATAH is percentage of time that active hosts experienced CPU utilization of 100%	The range is 0…100	SLA Violation Time per Active Host (SLATAH)	D	M	C					X				X	X				X			X							X	S25
PE	SLA violation	SLA violation(SLAV)	SLAV = SLATAH .PDM where PDM is Performance Degradation due to Migrations, SLA Violation Time per Active Host (SLATAH)		measures the SLA violation(SLAV)	D	M	C					X				X	X				X			X							X	S25
S	SLA/Security	SLA/Security			Compliance of SLA, security, privacy or copyright regulations	B	M	H	X	X			X				X		X		X	X	X					X					S36
PE	Video Quality	Sobel filter	[Sobel(Fn)] is each video frame (luminance plane) at time n(Fn) and it is first filtered with the Sobel filter .		measures the video frame (luminance plane)	D	M	C					X				X	X							X							X	S66
P	Software Tooling	Software Tooling	Sw		Software portability and API and SDK tools for developing cloud apps.	D	M	C		X	X	X	X				X	X	X		X	X						X					S36
PE	Operation time	Speed		PFlops, TPS, WIPS	Number of operations executed per second, (PFlops, TPS, WIPS, etc.)	B	M	C		X			X		X		X			X	X	X						X					S36
PE	Speedup	Speedup	Su		Speed gain of using more processing nodes over a single node	D	M	C		X			X				X	X		X	X	X						X					S36

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
M	Stability	Stability	$\sum (\alpha_{avg.i} - \alpha_{sla.i}) / (T/n)$ where α can be computational unit, network unit or storage unit of the resource, $\alpha_{avg.i}$ is the observed average performance of the user i who leased the Cloud service, $\alpha_{sla.i}$ is the promised values in the SLA, T is the service time, n is the total number of users In the case of a computing service, the stability is defined as the difference between the achieved performance and the specified performance in the SLA: St(t1, t2) = avg(t1, t2), where ς is a unit of computing, network or resource storage, ςu-avg is the average performance of the requested service between the time instant	time in miliseconds	The changeability or variability in the performance of a service. For computational resources, it is the deviation from the performance specified in SLAs, its performance variation in time. For storage service, it is the variation of the average read and write time.	D	M	C	X	X			X				X	X		X	X	X										X	,S24,S3
M	Stability	Stabilize rate	Error (t)= 1-λ(t)*r(t)	The range is 0 to 1. Err(t) function greater than 0 to stabilize the system to operate under target response time.	rate to stabilize the system to operate under target response time.	D	M	C					X				X	X			X	X										X	S31
M	Customizability	Static changes	Number of static changes in a Cloud service/workload (Ns)		measures the static changes in a Cloud service with respect to a workload	D	M	C					X				X	X			X											X	S65
R	Steady-state availability	Steady-state availability	$A(\infty) = \lim_{T \to \infty} A(T)$		The steady state availability can be determined by calculating the limit of the instantaneous availability as the time goes to infinity.	D	M	C					X				X			X	X	X									X		S52
PE	Storage capacity	Storage Capacity		size in MBs, GBs, TBs	measures the storage capacity with virtual disks to serve many user groups. Measured in GB), , Sg, , GBs	B	M	C	X	X	X		X			X	X			X	X	X	X	X				X			X		09,S10,S24,S36

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Storage horizontal scalability	Storage changes	Permissible storage changes to increased workloads	size in MBs, GBs, TBs	measures the permissible storage changes to increased workloads: GBs	B	M	C			X		X			X	X			X	X	X		X							X		S69
PE	Storage cost	Storage cost	$\sum C_j*D_j$ where Dj is the maximum data to be stored in the storage resource D sj associated to VM I, Cj Execution time of a task j on especific VM		measures the storage cost	D	M	C					X				X			X	X	X										X	S13
PE	Storage Capacity	Storage size		size in MBs, GBs, TBs	measure the size of storage data. GBs, Content --> Real size of storage data	B	M	C			X		X			X	X			X	X	X		X								X	48,S69
PE	StreamEye	StreamEye	SE	Byte/frame	Visualization of the video stream framing structure at the receiver end, probably impaired with frame losses	B	M	C		X			X				X			X	X	X	X					X					S23
FS	Image Degradation	Structural Similarity Index	SSIM	index/frame	Structural Similarity Index in Y component that measures the image degradation as perceived changes in structural information	B	M	C		X			X				X			X	X	X	X					X					S23
U	Success ratio	Success ratio	Success ratio : (no of successful operations in a workload) / (total operations available in the workload)	ratio, the range is 0...1	measures the ratio of successful operations	D	M	C					X				X	X			X											X	S65
FS	Suitability	Suitability of non-essential features	Suitability= number of non-essential features provided by service/ number of non-essential features required by the customer.	if only essential requirements are satisfied= 1 if all features are satisfied = 0 otherwise.	Suitability is defined as the degree to which a customer's requirements are met by a Cloud provider.	D	M	H		X							X			X		X									X		S24
R	Continuity	System Continuity	Rs(x) = Rs´(x)*Rs´´(x) where Rs(x) is system reliability for assignment X. Rs´(x) is system reliability without considering failure of links. Rs´´(x)is system reliability without considering failure of servers. X = [x ij]is task to precessor assignment matrix		measures system reliability for assignment x.	D	M	C					X				X			X	X	X									X		S22
PE	Connections	TCP connections		number of TCP connections of the server	measures the TCP connections	B	M	C			X	X	X				X			X	X	X										X	S35

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
M	Testing time	Testing time	Testing Time = Time to prepare test environment + Time to execute Test Suite for a Cloud workload. (Test Suite is collection of test cases).	time	measures testing time	D	M	C					X				X		X		X											X	S65
PE	TGuard	TGuard	TGuard: Trans-->ɸ(C)		clock guard	B	M	C					X				X			X	X											X	S48
PE	Throughput	Throughput		request/s	measures the rate at which requests are processed	B	A	C					X				X	X			X											X	S70
PE	Throughput	Throughput	the number of service requests served / total service time where a request can be jobs, tasks,operations	PFops, WIPS	measures the volume of requests per unit time (PFops, WIPS.)	D	M	C		X			X			X	X	X	X	X	X	X						X				X	,S49,S5
PE	Throughput	Throughput	measured by the time it takes to transmit a file of a certain length.		Measured as the average maximum data rate from sender to receiver when transmitting a file.	D	M	C					X				X	X			X										X		S54
PE	Throughput	Throughput	ß =n/B(n) where B(n) is the time needed to send n bytes	MBs/s	measures the number of bytes per second transmitted from the source to the destination across the network	D	M,A	C		X	X		X				X	X		X	X	X		X								X	,S37,S6
PE	Throughput	Throughput	It is defined by the maximum throughput or by the undergoing change of throughput with service intensity.		Is the service rate that a service provider can offer.	D	M	C					X				X	X	X	X	X	X									X		S78
PE	Throughput	Throughput	Throughput= (Tasks - Completed)/ (#Computations /Cloud) -Resource		number of tasks or works completed by the cloud service per unit of time	D	M,A	C	X	X			X				X	X		X	X	X			X						X		S26,S40,S6
PE	Throughput	Throughput App	Thrapp=n/(T-To) T: the execution of tasks of an application in traditional data center To: the overhead of Cloud data centre n:collection of tasks		measures the number of tasks completed in unit time by the cloud service. Throughput of an application not only depends on service provider but also on application itself.	D	M	C	X	X			X				X			X	X	X	X								X		S09,S10

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Throughput	Throughput Cloud	total throughput of a Cloud service α =n/Te(n, m) + To Te(n, m) be the execution time of n tasks on m machines. To be the time overhead due to various factors such as infrastructure initiation delays and inter task communication delays Let an user application have ‘n’ tasks and they are submitted to run on ‘m’ machines from the Cloud provider.		Throughput is the number of tasks completed by the Cloud service per unit of time.	D	M	C		X							X			X		X									X		S24
PE	Time constraint	Time constraint	Time=(C,TGuard,CReset) where C is the clock variable, Tguard is clock guard, Creset is reseting clock		measures a time constraint	D	M	C					X				X			X	X											X	S48
PE	Time Behaviour	Time Behaviour	Behavior of Time (BoT) BoT =waiting time/Total time for service invocation The denominator is the time taken for sending the requests and receiving a response. The numerator is waiting time for the service to get executed	The range is 0...1 and lower value indicates that the SaaS has higher efficiency of time.	measures a ratio of waiting time for the total time for invocation.	D	M	C					X				X	X			X										X		S53

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Time Behaviour	Time Behaviour	TB =(executiontime)/(total service invocation time) where the denominator is the consumed time period from sending request to receiving response. The numerator is an execution time just for processing functionality.The execution time can be computed as (total service invocation time – waiting time).	The range is 0...1 and higher value indicates that the SaaS has higher efficiency of time.	measures a ratio of an execution time for a total invocation time.	D	M	C		X							X	X			X										X		S44
R	Mean Time between Failures	Time between consecutive service failures	NormalOperationalPeriodDuration/NumberOfFailures		measures the Time between consecutive service failures	D	M	C			X		X			X	X	X		X	X	X		X							X		S69
PE	Time slot	Time slot utility function	time slot utility UPT(X) = uX,min + (1 + uX,min)* $\vert$RXp -X / RXp-IXp$\vert$; IXp ≤ X ≤ RXp, otherwise 0 u T,min is the minimum utility that the provider receives for reaching a deal at its worst time (or least preferred) slot. Let IXC and RXC (IXp and RXp) be the most preferred (initial) service quality and the least preferred (reserve) service quality for a consumer agent (a provider agent), X is the service quality to be evaluated	ratio, The range is 0...100. consumers prefer the highest guarantee for a service quality X, and providers want to sell their services with the lowest guarantee for the service quality with a given price	designed to model consumers' and providers' preferences for different time slots. Provider's time slot utility function Providers prefer to allocate jobs: 1) to their earliest available time slots, and 2) to the time slots at which the job sizes can be accommodated to optimize resource utilization. Consumer's time slot utility function When reserving a Cloud service, a consumer generally describes diverse forms of temporal preferences with multiple sets of time slots.	D	M	C		X							X	X		X	X	X									X		S67
PE	Execution time	Time Spent with Physics (TSP)	TSP= tPhysicsSimulationResults – tSimulatePhysics.	Time in minutes	measures the time spent with physics	D	M	C			X		X			X		X	X	X	X			X							X		S32
PE	Execution time	Time Spent with Rendering (TSR)	TSR = tSceneDataUpdated – tUpdateSceneData.	Time in minutes	measures the time spent with rendering	D	M	C			X		X			X		X	X	X	X			X							X		S32

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Timeliness	Timeliness	TL(i) = Twait(i) +Texec(i) +Ttrans(i) TL(i) represents the total completion time, Twait(i) is the waiting time, Texec(i) is the task execution time, Ttrans(i) indicates the transmission time of the i-th task during VM migration	time in milliseconds	Timeliness is an ability with which a cloud monitoring system is able to supply information in time when users need to access it. measures the total time taken to complete a task. measured as delay, or latency, or time to complete.	D	M	C	X	X			X				X			X	X	X									X		S26
PE	Server Capacity	Total processes number		Total processes number	measures the total number of process	B	M	C					X				X			X	X	X										X	S35
PE	Jitter	Transcoding Jitter Probability	ψ = Tj/Ttotal, where Tj specifies the total jitter time and Ttotal means the total runtime of the system	a higher transcoding jitter probability may lead to more transcoding jitters	indicates the probability that the TRB exceeds the MOB. Where MOB is considered as the total transcoding capacity and TRB is the transcoding demand and the metric of MOB & TRB is bitrate	D	M	C					X				X	X			X										X		S57
PE	Turnaround Time	Turnaround efficiency	TE = Turnaround Efficiency for a job by resource Rk = Promised Turnaround time by Rk in the SLA / Actual Turnaround time by Rk to complete the job		Turnaround efficiency of a resource Rk (TE) is the average of turnaround efficiency . Turnaround efficiency is 1 if the promised turnaround time is greater than the actual turnaround time	D	M	C		X							X	X	X	X	X	X									X		S50
PE	Turnaround Efficiency	Turnaround efficiency (TE)	TE=w*AV + w*RE+w*DI+w*RT, where AV is availability, RE is reliability, DI is data integrity, RT is response time		measures turnaround efficiency (TE)	D	M	C					X				X	X	X	X		X										X	S27
PE	Packets received	tx_bytes		Bytes sent	measures the bytes sent in Packets received	B	A	C					X			X	X		X		X											X	S70
PE	Packets sent	tx_packets		number of packets	measures the packets sent (tx_packets)	B	A	C					X			X	X		X		X											X	S70
P	Underprovisioning	Underprovisioning ratio	UDPR DT = 1 - 1/n n as a number of the requests received during a period of time DT and i the number of the requests where the requested resources are beyond the available resources	Range 0..1. Value 1 implies that no request in under-provisioned. Value 0 means that all the request are under-provisioned	Under-provisioning measures the ratio of the number of the unsatisfied request increases.	D	M	C		X							X			X					X							X	S02
P	Underprovisioning	Underprovisioning resource	∑ underprovisioning resource		measures provisioning resource is under the needs, the SLA and QoS are not optima	D	M	C					X				X			X	X	X										X	S35

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: TheoreticalValidation		Empirical Validation			No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
M	Readability	Understandability of Service (UoS)	Understandability of Service (UoS) UoS = 1 - No. of fields with unacceptable readablity / total number of fields		measures the ratio of amount of fields which are unacceptable readability to the total number of fields	D	M	C					X				X	X			X										X		S53
R	Service Availability	Availability	Availability=MTBF/(MTBF + MTTR) where MTBF represents the mean time between failures (uptime) MTTR represents the mean time to repair (downtime)	percentage The higher the uptime percentage is closer to 1	measures the ability of a CS Users to access the cloud resources. The ratio of the time a service component is functional to the total time	D	M,A	C	X	X			X				X	X	X	X	X	X	X									X	,S65,S7
R	Service Availability	Uptime percentage	Av =∫ 1/fdp(t) * dt	percentage The higher the uptime percentage is closer to 1	Av, is the percentage of time during which the user can access the service,	D	M	C		X			X				X	X			X	X									X		S30
R	Service Availability	Uptime percentage	Av(s)=100% -UAv(s) UAv(s)= (ΣSF(t) / AG(t)) x 100% SF(t) is the service unavailable time during the negotiated time, AG(t) is the negotiated time.		Av percentage of time during which services can be used at the quality level specified in the SLA contract. UAv(s) is the percentage of service unavailable time during the negotiated time.	D	A	C					X				X	X	X	X	X	X									X		S08
R	Service Availability	Uptime ratio	Availability = ((total service time) –(total time for which service was not available))/ (total service time)	percentage The higher the uptime percentage is closer to 1	percentage of time for accessing the cloud service,and customers can access it. Be expressed in terms of immediate use or average downtime per week, month or year or as total downtime for a given week, month or year	D	M,A	C	X	X			X				X	X		X	X	X										X	5,S24,S
R	Service Availability	Uptime ratio	Availability = TotalUptime/TotalTime	percentage The higher the uptime percentage is closer to 1	The availability is the percentage of uptime duration over a defined interval duration. availability means that the information of a cloud computing system is accessible to those who have been authorized to have access at appropriate times.	D	M,A	C	X	X	X		X			X	X	X	X	X	X	X		X	X			X			X	X	3,S83,S(
U	Usability	Usability	Usability = Previous user experienced the cloud service/average time		simplicity of using a cloud service. average time qualified by the preceding customers of the Cloud service to activate, learn, install and understand it.	D	M	C	X	X			X				X			X	X	X										X	S07

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
U	Usability	Usability	This is the ease with which a service can be used, operated, learned, understood, and installed by the user	{High, Medium, Low}	measures the ease with which a service can be used, operated, learned, understood, and installed by the user	D	M	Q		X			X				X	X			X	X										X	S21
U	User interfaces	User interfaces		list of user interfaces available: GUI, WUI, API graphical, web or application	Usability (USAB) describes how easy, efficient, and enjoyable the interface to a cloud service is to use, or assesses the ease of invocation if the functionality of a cloud service is exposed as Application Programming Interface (API)	B	M	Q			X		X				X	X	X	X	X			X	X			X	X			X	S82,S83
PE	Service use rate	User Self Service Rate	Self Service Rate =100%-(# of Inquiries regarding workloads/# of Support Resource Visits)		measures the user self service rate	D	M	C					X				X		X		X											X	S65
PE	Performance utility	Utility of performance for request(Pi)	Pi=1/(wt max ϵ {tij} + wc ∑ cij+ we∑ eij) T is waiting time matrix, C is cost matrix and E is energy consumption matrix. where wt, wc, we are weights of waiting time, cost and energy consumption. Sum of weights is equal to 1.		measures the utility of performance for request determinated by waiting time, cost, energy consumption.	D	M	C					X				X			X	X	X									X		S75
PE	Data center utilization	Data center utilization	U = min{ 1;E [ru x] /N } Virtual resources		The expected data center utilization U can be computed as the ratio between the number of physical resources used at steady state and the total number N of physical resources.	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Resource Utilization	Time Utilization	Utilization= (time that cloud resources are utilized/ total time of cloud resources) *100		measures the percentage of time that cloud resources are utilized	D	M	C					X				X	X	X	X	X	X									X		S78
M	Variability	Variability	Variability= standard deviation of the metrics	should be as low as possible to ensure the consistency of these provided services	measures discovered over time using the standard deviation of the metrics	D	A	C					X				X	X	X	X	X	X									X		S61

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure: Theoretical Validation		Validation procedure: Empirical Validation			Validation procedure: No Validation		Ref. Paper
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Performance of Virtual Machine	Variability through relative standard deviation (RSD)	RSD = 1/E(x)√ 1/(N-1) * ∑(xi − E(x))^2, where N is the total number of measurements, x1, ... , xN are the measured results, and E(x) is the mean of those measurements.	The range is 0 to 100 lower variability indicates a lower variation in the taken performances measurements	measures the variation in the taken performances measurements for each benchmark. RSD is calculated from the data collected through all the measurements done for one metric.	D	M	C		X							X			X		X										X	S81
PE	Variable Computing Load	Variable Computing Load	Variable Computing Load = Change in Load Balancing (ΔLB) ΔLB= Actual load at time t/Expected load at time t	ΔLB ≤ 1 for efficient Cloud service	measures the change in Load Balancing(ΔLB)	D	M	C					X				X	X			X											X	S65
P	Vertical scalability	Vertical scalability	vertical scalability=Σmi Σnij (proportion of increase in rij). where rij be resource j that needs to be enhanced on Cloud service i, n and m be the number of resources assigned to a particular Cloud service and the number of Cloud services used by the user, respectively.		The vertical scalability can be calculated as the maximum available increase in the resources of a Cloud service.	D	M	C		X							X			X		X									X		S24
PE	Performance for video streaming	Video Quality Score(VQS)	VQS = Σ Qi * Duration/Qmax ∗ Total Duration ∗ 100 where Qi quality displayed, Qmax maximal quality that can be displayed	resolution	to quantify the overall display quality (resolution) during a streamed video viewing session taking into account the duration of each quality displayed and the maximal quality that can be displayed by the device subject to its screen size	D	M	C					X				X	X							X						X		S64
P	Virtual Systems resources	Virtual Systems		Virtual Systems. SaaS a locally-hosted application PaaS OS and associated services IaaS involves outsorcing equipment, storage, hw, servers, networking	identify the cloud virtual infrastructure	B	M	C	X			X	X			X	X	X	X	X	X	X									X		S26

Taxonomy of metrics for cloud services

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
P	Virtualization platform	Virtualization measure		Virtualization platform. List of virtualizations	identify the Virtualization platform support	B	M	Q	X	X	X	X	X		X		X	X	X	X	X	X	X	X							X		S10
R	VM availability	VM available probability	pa = πG + πA = 1 − πF Sum i∈(G,A,F) πi = 1. steady-state probabilities of states: G good, A attack, F fail.		defines the probability that there are x VMs available among n VMs. pa represents the probability that the VM is available.	D	M	C					X				X			X	X				X							X	S79
PE	VM Energy Consumption	VM Communication Energy consumption	energy consumption for communication among all VMs in the server S, ∑EVCom= ∑ ∑dij/bS·wS where dij represents the amount of data transferred from Vi to Vj in S, bS is the internal communication bandwidth of S, wS is the energy consumption rate for the internal communication among VMs of S.	Energy consumption	measures the VM Communication Energy consumption	D	M	C					X				X	X							X							X	S19
PE	VM Energy Efficiency	VM Energy efficiency	EfVMj	Energy consumption	measures the energy efficiency at VM level for IaaS providers. the energy efficiency of a VM which is running in node i	D	M	C					X				X			X		X										X	S41
PE	VM Energy Efficiency	VM Iddle Energy Consumption	∑EIdle = ∑tj · Pl = ∑ tj ·ξ S i-α where l is the number of iddle VMs in the server; Pl is the power of the iddle VMs, Pl =ξ Si-α , Si is the iddle speed of VMs in the server; tj is the iddle time of the j-th VM in S	Energy consumption	measures the VM Iddle Energy Consumption	D	M	C					X				X	X							X							X	S19
PE	VM Migrations	VM Migrations		number of VM Migrations	measures how much migrations happened during VM allocation	B	M	C					X				X			X	X	X										X	S01

QoS Characteristic	Attribute	Metric	Function	Result	Description	Metric Type	Tool Support	Measurement result	Lifecycle phase						Artifact			Service Type			Stakeholder viewpoint					Validation procedure							Ref. Paper
																										TheoreticalValidation		Empirical Validation			No Validation		
									Req.	Adq.	Dev.	Int.	Ope.	Ret.	Spe.	Arc.	Ser.	SaaS	PaaS	IaaS	CSP	CSC	CSB	CSD	USE	A.A.	A.T.M	C.E.	C.S.	SU.	N.V.	P.C.	
PE	Energy Efficiency	VM Run Energy Consumption	∑ERun = ∑ti · PR = ∑ ti ·ξ S r-α where R is the number of running VMs in the server; PR is the power of the running VMs, PR =ξ Sr-α , Sr is the running speed of VMs in S; ti is the execution time of the i-th VM in S	Energy consumption	measures VM Run Energy Consumption	D	M	C					X				X	X							X							X	S19
PE	Waiting Time	Waiting time	Ri=max {tij| tij ϵ ti}	time in milliseconds	measures the waiting time of request. Tij represents the amount of time when aij data block have been prepared for request Ri fron VMj.	D	M	C					X				X			X	X	X									X		S75
PE	Waiting Time	Waiting time	W = E[rq1x] + E[rq2x] /Thr(Tarr) * A where Thr(Tarr) is the expected throughput of transition Tarr. rq1 = P# queue; rq2 = P# send;	time in milliseconds	Waiting Time measures the expected time W spent by users in the queue (either the system or the upload queue).	D	M	C		X	X						X			X	X	X		X								X	S14
PE	Web server access	Web access rate		number of accesses/second	measures the number of accesses per second	B	M	C					X				X		X		X											X	S65
PE	Workload rate	Workload arrival rate	λ(t)=λ(t-1)/λmedian		current arrival rate in the system is divided by median of previous arrival rates.	D	M	C					X				X	X			X	X										X	S31
PE	Latency	writeLatency		time in milliseconds	measures the Keyspace write latency	B	A	C					X				X	X			X											X	S70