

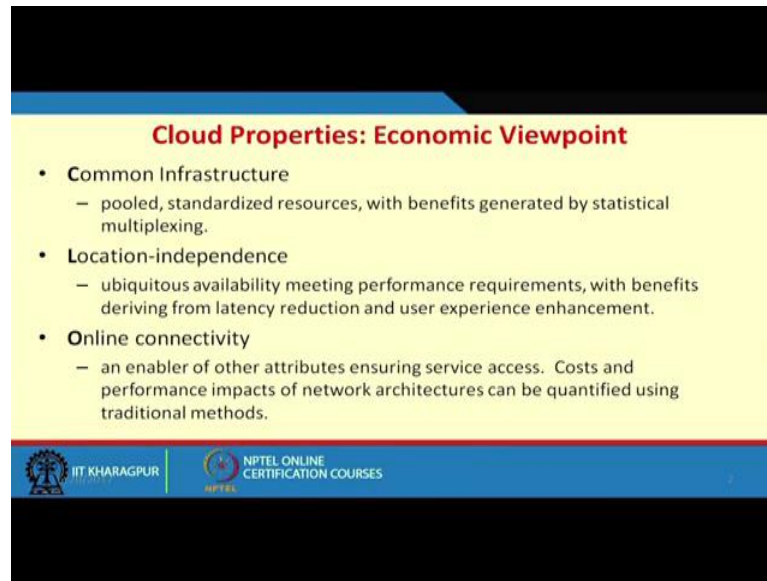
Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Lecture – 12
Economics

Hello. So, we will continue our discussion or lectures on cloud computing. Today we will take up a topic, which is we need to look at that what is this economy behind cloud computing. Why people will go for this cloud computing type of things right. What it is not like that, we are getting a something totally new. So, it is only that the different type of application etc; now we are getting as a service. So, it is what makes this or what will make this viable. Is it always going to cloud is beneficial or where how to decide that; whether I need to go to cloud or whether I need to buy something in house and how much how to balance between in house infrastructure and provisioning of the cloud. So, in order to do that, we will try to see some basic phenomena or basic economic point which is makes cloud viable at which type of operations or which type of situation type of things.

So, we try to more try to look out, that when we organization or individual especially organizations, one switch over cloud partially or fully, what are the consideration it should keep in mind. We have seen SLAs they are there are other issues that even during cloud our initial lectures that there are some limitations, there are some issues with the cloud, but keeping all that even all those things are running fine. Whether it is economically to be on cloud always or at times or at what times and type of things and what should be the business consideration. So, we will have a brief discussion on that. So, those of you are interested in working on this line can basically now leverage on this type of work.

(Refer Slide Time: 02:20)



Cloud Properties: Economic Viewpoint

- **Common Infrastructure**
 - pooled, standardized resources, with benefits generated by statistical multiplexing.
- **Location-independence**
 - ubiquitous availability meeting performance requirements, with benefits deriving from latency reduction and user experience enhancement.
- **Online connectivity**
 - an enabler of other attributes ensuring service access. Costs and performance impacts of network architectures can be quantified using traditional methods.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, from the economic point of view, if we see that what are the different cloud properties relook at the properties one is the common infrastructure.

So, pooled standardized resources with benefits generated by statistical multiplexing right? That is important. First of all, it is a pooled standardized resources. So, at the service provider end it is a pools of resources, which are standardized how they can be integrated, etcetera, right. What is the benefit from the thing because of multiplexing tasks, right. So, statistically multiplexing benefits statistical multiplexing; that means, I have some 100 systems, right. So, it is highly underutilized even 15-20 percents are not utilized of my own workload, or from some workload. So, if I have multiple workloads, I may with the same type of system. I can go up into the things, what I am thinking that why I have purchased this some 10 or 100 systems. Because considering most of the cases, we have considering a peak workload, right.

At peak time I will be require all those things right. Say for example, IIT kharagpur in a particular department, a particular lab for a particular year, say MTech first year. The numbers of seats are say 50 for a particular department and what we expect that 50 percent we will do work on 50 individual systems; we purchase a lab we basically set up a lab of 50 systems. Adjoining we same time we have to go for power, at the same time we have to go for AC and we feel that if there is a system goes out and type of things, keep another 5, 10 percent another 5 system on a standby. So, having 55 systems, but it

may so, happen that the recruitment the number of students joined the course may be less than 50. So, I have any surplus power even they joined the courses right. This 50 system are again underutilized, not may not be utilized all the things during the lab hours. May be daily 8 hours it is utilized otherwise things are not there, right. One is looking at the peak load.

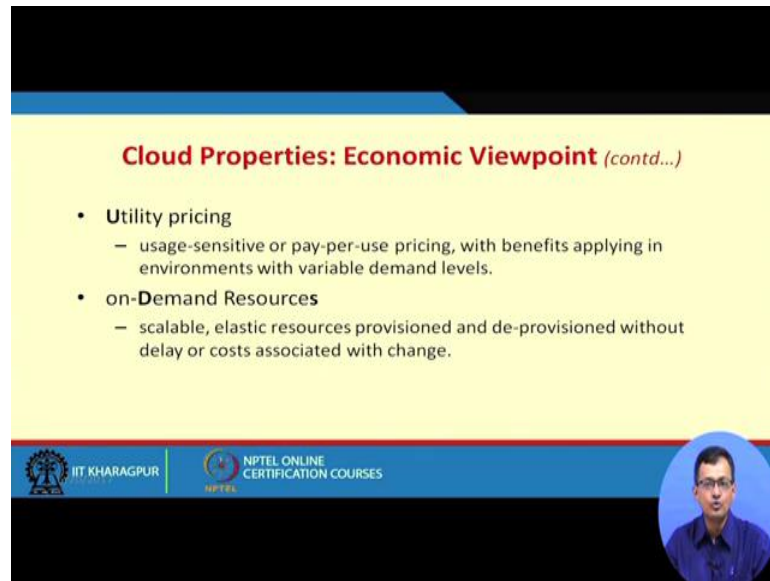
So, but whenever I consider the system individually. I have taken this processor memory etcetera thinking, this lab assignment will be leading up to that level so; that means, at the peak load, right. So, this consideration may be or at many times. I do lot of over sizing of the things.

And if I have is it becomes more critical, when I have a server, a single server which is catering to number of users and then I think that all the user will jump into the thing right and then the peak load will be there right. If those who are working in the networking you might have seen that typically 10, 24 port networks switch right. How many people can connect? 24. Even 100 mbps line it is 24 roughly, 24x100 if we not go to the integrity of this binary thing.

So, it is approximately 2.4 gigabytes, but the switch uplink maybe 1 gigabyte. So, it is some sort of a blocking architecture, but if I give a uplink of 2.4 or 3 gigabyte then it is over provisioning, right. I it is all the people are coming statistically at the same peak time may not be there. So, it statistically we need to look at that whether it is viable to provision. Such a higher thing when you provision higher, thing it involves lot of costing and other things in to the things maintenance of the thing, there are things of adjoining other accessories, etcetera.

Even the cost of the equipment goes up so and so far. So, that is one thing, another is the location independence that is another property of cloud like ubiquitously available meeting performance requirement with benefits deriving from latency reduction and user experience in enhancement. So, this is a location independence property. So, whether it is economically, how to make use of that online connectivity as an enabler of other attributes ensuring a service cost and performance impact on the network architecture etcetera. So, I should have online always connectivity. So, there is another factor. So, these are the different factors which may force us to look at the different things.


(Refer Slide Time: 06:55)



Cloud Properties: Economic Viewpoint (contd...)

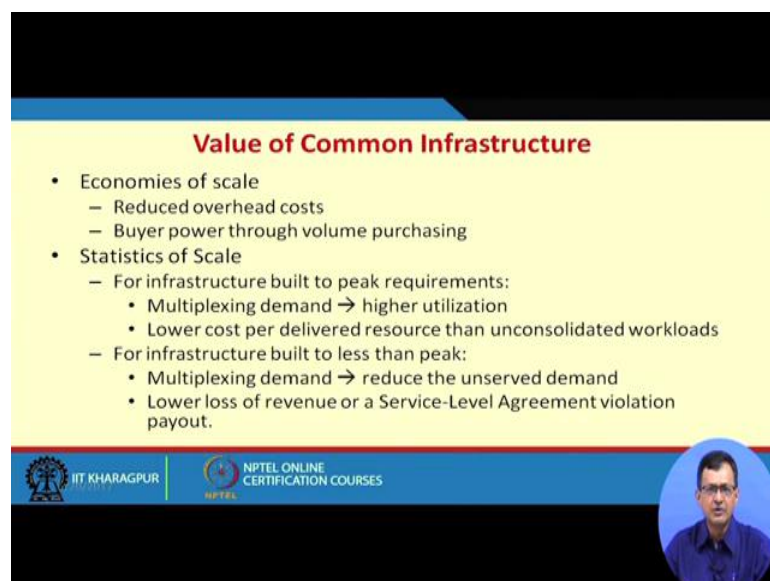
- **Utility pricing**
 - usage-sensitive or pay-per-use pricing, with benefits applying in environments with variable demand levels.
- **on-Demand Resources**
 - scalable, elastic resources provisioned and de-provisioned without delay or costs associated with change.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There are other two direct economic factors, one is the utility pricing right. I pay as you go model, I pay per unit things like electricity etcetera. There is another thing called on-demand resources. So, when I demand for the resources are provisioned right. So, it is on demand for the resources are provided. So, scalable, elastic resources, provision and de-provision without delay and cost associated with the change, right. So, there cannot be delay or cost associated anything. There may be cost factor, rather there should not there minimal human on managing managerial involvement. So, that as I go and go forth and things. I may resources provisioning and de-provisioning.


(Refer Slide Time: 07:43)



Value of Common Infrastructure

- **Economies of scale**
 - Reduced overhead costs
 - Buyer power through volume purchasing
- **Statistics of Scale**
 - For infrastructure built to peak requirements:
 - Multiplexing demand → higher utilization
 - Lower cost per delivered resource than unconsolidated workloads
 - For infrastructure built to less than peak:
 - Multiplexing demand → reduce the unserved demand
 - Lower loss of revenue or a Service-Level Agreement violation payout.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



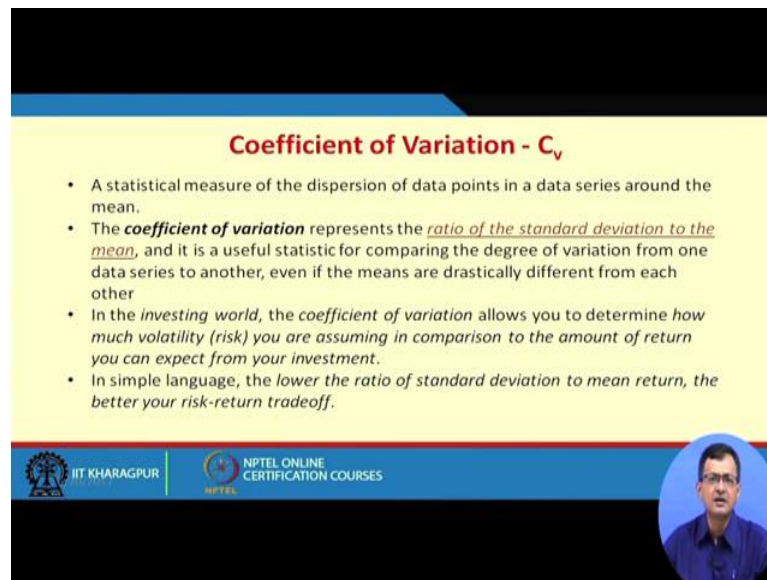
So, with this we want to try to find out this valuation of this or whether there is an economic point of view. We try to look at a little bit statistically, very little portion of them need to understand this issue. One is economies of scale, reduced overhead cost buyer power through volume purchase right. One is that, I want to one is the economy scale means I want to reduce this overhead cost right. So, what are the different overhead costs like if I buy a system, it has overhead cost of AC, it has overhead cost of power maintenance like UPS or it has overhead cost of AMC of the system right? And there are several others of like human resources to maintain the thing etcetera. So, I want to reduce that overhead cost, right.

So, buying the thing is still, but maintaining the thing at times becomes over the years much costlier than the equipment itself right. And there is a major problem of especially for in the computing world that after typically 2 to 3 years. Not more than definitely within 5 years, that whole thing becomes obsolete. The technology becomes obsolete the whole system power etc, is no longer valid becomes viable for installing new set of tools and software etcetera. So, that is a big problem. There is a statistics of scale on the other end. For infrastructure build on peak requirement like infrastructure, build up peak requirement.

I think that, I build an infrastructure that always all my students will be there in the class or all will register for the course and etcetera. So, multiplexing demand may help me higher utilization right. So, when I use things on multiplexing different demand may help me higher iteration. Lower cost for delivered resources than unconsolidated work. So, it is a lower cost per delivered resources than if there is an unconsolidated workload. So, if it is a consolidated and lower cost will be there.

So, for infrastructure build to less than peak. So, it is not peak less than peak, here also multiplexing demand reduce the unmet demands right. So, multiplexing there may be if it is a blocking architecture, there can be some of the things which are not served. So, multiplexing may reduce this unmet, this it is not like that it is good that one it is always based to the one. I go on some sort of a scheduling algorithm and go on serving. Lower loss of revenue or a service level agreement violation. Because, SLA violation means you need to pay out something SLA violation payout. So, it may reduce lower loss of revenue and etcetera. So, both for peak infrastructure build on peak and non peak we have this sort of things.


(Refer Slide Time: 10:30)



Coefficient of Variation - C_v

- A statistical measure of the dispersion of data points in a data series around the mean.
- The **coefficient of variation** represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other
- In the *investing world*, the *coefficient of variation* allows you to determine *how much volatility (risk) you are assuming in comparison to the amount of return you can expect from your investment*.
- In simple language, the *lower the ratio of standard deviation to mean return, the better your risk-return tradeoff*.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Another term which comes not only for cloud, for any type of these things where this key is involved and you have to give services what you say coefficient of variance or commonly CV.

So, it is not exactly covariance we are talking about, it is coefficient of variance. So, a statistical measure of the dispersion of data in a data series around the mean, right; like a coefficient of variance is represented by the ratio, of the standard deviation of the mean to the mean. So, it is standard deviation or those who understand sigma to mu sigma by mu right. Here standard deviation by mean and it is useful statistics for comparing the degree of variation from one data series to another right. So, I can say that, this whether the coefficient of variation of this data series, is more than less than equal to other data series is this is a good measure to look at it, right, rather even if the means are drastically different then also we can compare to data series.

So, how these two data series behavioral things are there the CV gives me idea to do that right. So, it is widely used or widely looked into in the investing world. So, in investing world coefficient of variation allows, you to determine how much volatile or risk, you are assuming in comparison to the amount of return you can expect from your investment.

So, this CV gives you idea that, how much risk you are assuming in comparison to the amount of return you can expect from the investment. So, how much risk you are

involving. So, this is important if you see, it is also important in our sort of scenarios also like we are we are basically involving some risk by leverage, by putting my organizational from means infrastructure or on from the on premise infrastructure. So, cloud infrastructure. So, there is a risk of that if the infrastructure is not available so and so far. So, it is not only infrastructure I want to mean to say that, all type of services on the cloud thing. So, if the service is not available at it. So, how much risk I need to take on those things. So, in simple language lower the ratio of standard deviation to mean the better is your risk return trade off right.

So, lower the ratio to standard deviation to mean return. So, the better is the risk. So, I can have more smoothness into the curve that is important.

(Refer Slide Time: 13:05)

Measure of "Smoothness"

- Coefficient of variation C_v
 - \neq the variance σ^2 nor the correlation coefficient
- Ratio of the standard deviation σ to the absolute value of the mean $|\mu|$
- "Smoother" curves:
 - large mean for a given standard deviation
 - or smaller standard deviation for a given mean
- Importance of *smoothness*:
 - a facility with fixed assets servicing highly variable demand will achieve lower utilization than a similar one servicing relatively smooth demand.
- **Multiplexing demand from multiple sources may reduce the coefficient of variation C_v**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is some sort of a measure of smoothness. So, if it is a very variable load or lot of peak and non peak things, then I am in much bigger trouble in measuring that how things will be there. If it is a smoothing out load then, I am much in a better position to do that. So, coefficient of variation CV as you says that, it is not similar as variance, nor correlation coefficient as we are mentioning ratio of standard deviation to sigma, to absolute value of the mean mu or mu mod.

So, smoother curves large mean for a given standard deviation. So, as we are telling the sigma by mu, if it is a large mean then given standard deviation things will be not varying that much or a smaller standard deviation or of a given mean that also things will

be much under control. So, importance of smoothness, a facility which fixed asset servicing highly variable demand will achieve lower utilization than a similar one servicing relatively smooth demand right. I have I am a service provider, I have some 100 systems are my backbone and I know that the demand will be something smooth right.

Then I can basically have a better management of the things or I have n number of systems. So, what will be the value of n is much easier, but if it is a very much varying suddenly demand goes up 100 then 10, etcetera, then I have a problem, right. Similarly for any, if you look if we look at our day to day life any shop shopkeeper, the amount of provisioning you will do that you will keep this in it is store, this depends on the demand thing, if the demand is something smooth, it may vary that Monday demand is different from Tuesday demand, different from week weekdays to weekends demand etcetera. Fine, but he has a idea, but it is totally random or you done cannot do anything, then it is a very difficult to provisional things. So, that is the same thing holds there.

So, multiplexing demand from multiple sources may reduce the coefficient of variation right that is the thing. So, now, I cannot predict that who what are the different sources how things will be there, but if I could have multiplexed the different demand from multiple sources. So, it may happen that this multiplex thing may have a give me a better CV right. Which may have a little smooth CV, where the overall demand things are there it is not like that all are going peak at the time, all are coming down at the lower things, but I have a multiplex type of things. Like if we look at say X_1, X_2, \dots, X_n are independent random variables of demand identical standard, say for our argument sake they are having though they are independent and then, but they have something identical sigma and mu.

So, aggregate demand in case of mean, some of the means $n \cdot \mu$ aggregate variance is $n \cdot \sigma^2$. So, if we calculate that coefficient of variance, it is $(1/\sqrt{n})C_v$, right. So, if n increases I multiplex more than the $1/\sqrt{n}$ decreases. So, I have more smoothing out, this coefficient of variation or moves smooth type of curve. So, adding n independent demand reduces C_v by $1/\sqrt{n}$. So, it is it becomes more smoothing out. So, penalty of insufficient excess resources go smaller, right.

So, what is happening if it is smoothing out then, I do not have to plenty I have to keep more resources at my back end right, it reduces right. Otherwise I do not know whether it is a 10 demand of 10 units or 100 units and etcetera and then I have to keep a track of 100 units at the back end. Like aggregating 100 workload bring the penalty down by 10 percent, 1 by root over root over 100 is 10. So, it may bring down the whole thing by 10 percent. So, aggregating multiple resources may allow me to have a reduced loading.

(Refer Slide Time: 17:23)

But What about Workloads?

- Negative correlation demands
 - X and $1-X$ Sum is random variable 1
 - Appropriate selection of customer segments
- Perfectly correlated demands
 - Aggregated demand : $n.X$, variance of sum: $n^2\sigma^2(X)$
 - Mean: $n.\mu$, standard deviation: $n.\sigma(X)$
 - Coefficient of Variance remains constant
- Simultaneous peaks

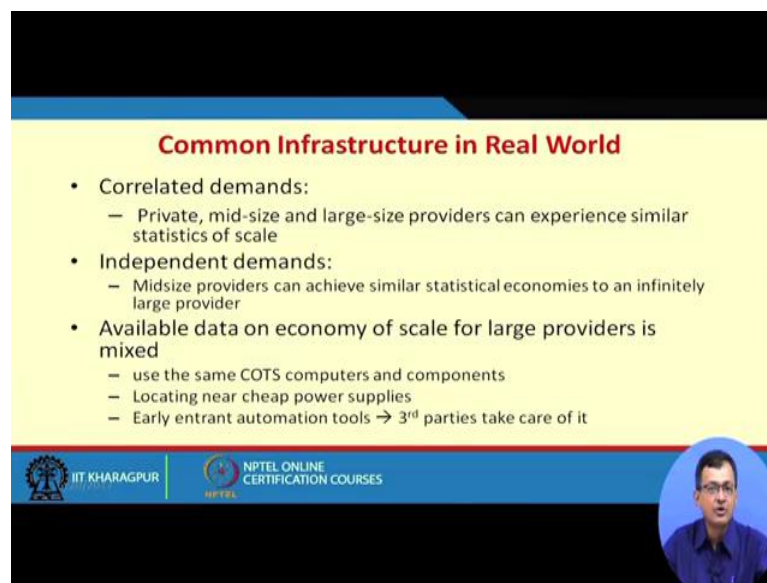
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

But, what about the different workloads; non negative correlation demands like if x and $1-x$ sum of a random variable is 1 appropriate selection of the customer segments right is another important thing right, I say that I have a computing infrastructure. I select those customers, some of them are active on day time, some of them processing are active on the night time, right.

So, it is compensating, right, I if the both are working at the same time then the peak will go much higher, but selecting that negative demands or the time in the timescale. I could have managed those thing with the same infrastructure. So, that is selection of the customer base is the important things, what sort of things will be there when I go on selecting as customers. Perfectly correlated demand, if some of the things are perfectly coordinated the aggregate demand will be $n.x$. Variations will be $n^2\sigma^2(x)$ and mean will be $n.\sigma$ and standard deviation $n.\sigma(x)$ and so far, coefficient of variation remains constant. So, it is perfectly correlated demand that this thing will be as a constant.

There are third issue if all demands are coming at the peak at the same time all at the peak at the same time then, I have a serious problem right. Then I have congestion at the things all are demanding at the same time. Like if I say all classes are breaking at the hourly basis. So, at the hourly basis, we have a huge demand for this root network, right. Lot of vehicles like said studying for cycles to it said they are on the demand because all classes are things. So, it is a peak coming to the thing at the same time like this then we have a problem.


(Refer Slide Time: 19:13)



Common Infrastructure in Real World

- Correlated demands:
 - Private, mid-size and large-size providers can experience similar statistics of scale
- Independent demands:
 - Midsize providers can achieve similar statistical economies to an infinitely large provider
- Available data on economy of scale for large providers is mixed
 - use the same COTS computers and components
 - Locating near cheap power supplies
 - Early entrant automation tools → 3rd parties take care of it

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES





So, common infrastructure in our real world, correlated demands, private, mid sides, large sized providers can experience similar statistic scales. So, it is a more or less correlated demand. Independent demand, midsize provider can achieve similar statistical economy to an infinitely large provider right. Available data on economy of scale for large provider is mixed right because use of same COTS type of things that is commercial of the self systems and components locating near cheap power supplies that is one thing like if as the power supply is a major thing for the data centers, they want to build a data center where the power supply will be much cheaper and nearby. Early entrain automation tool third party parties takes care of it. So, there can be early entrant automation tools which the third party can be taking.


(Refer Slide Time: 20:12)

Value of Location Independence

- We used to go to the computers, but applications, services and contents now come to us!
 - Through networks: Wired, wireless, satellite, etc.
- But what about latency?
 - Human response latency: 10s to 100s milliseconds
 - Latency is correlated with:
 - **Distance (Strongly)**
 - Routing algorithms of routers and switches (second order effects)
 - Speed of light in fiber: only 124 miles per millisecond
 - If the Google word suggestion took 2 seconds ☹
 - VOIP with latency of 200ms or more ☹


IIT KHARAGPUR


NPTEL ONLINE
CERTIFICATION COURSES



Say value of there is a value of corresponding to the location or value of location independence. We use to go to the computers, but the application services contains now come to us right. So, there is a paradigm shift like say yesterdays or we used to go to the system to work on it. Now the system power etcetera coming to my own desktop like; so, I have a large pool of huge resources on my very thin system, it can be a simple desktop, laptop and type of things all are provisions here, through networks wired wireless satellite etc, but what about latency?

So, latency also a big thing, right. So, human response is 10 to milliseconds, 10 to 100 milliseconds. So, latency is correlated strongly with the distance right more that distance, more the network latency another type of latency more on the hopes, more on the failure rates and etcetera.

So, though it also depends on what sort of routing algorithms etcetera also coming into play. So, speed of anyway we know that speed of light in fiber. So, some particular 124 miles per milliseconds. So, if suppose I am searching something and it takes more than couple of seconds, then we are not happy with that even a VOIP thing, if it is something delayed more than something, 200 nano second; a 200 millisecond second, then it is very difficult to communicate over this voice over IP.

(Refer Slide Time: 21:49)



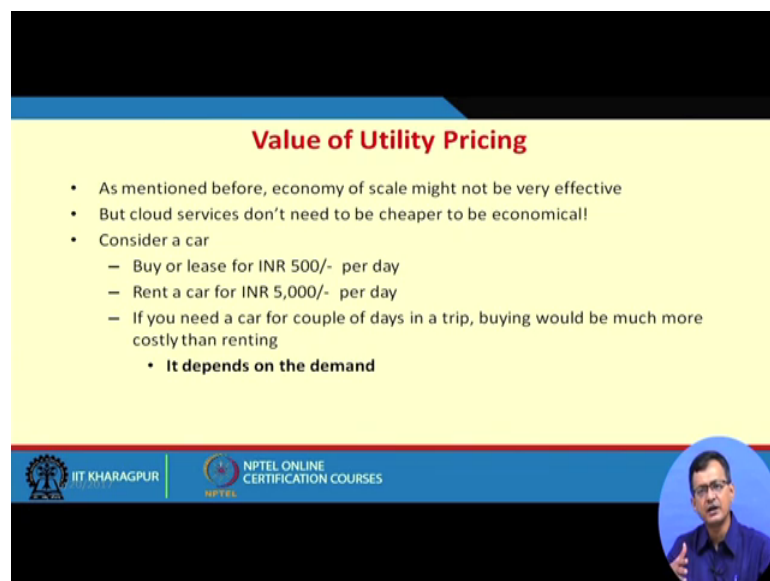
Value of Location Independence (contd...)

- Supporting a global user base requires a dispersed service architecture
 - Coordination, consistency, availability, partition-tolerance
 - **Investment implications**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there is a supporting of a global user base requires a dispersed service architecture. So, the architecture if I want to support, if my provider is to support all the global user base then, I have to have a appropriate distributed and dispersed architecture to do that and similarly the protocol. So, coordinate and coordination consistency availability partition tolerance these are issues. So, and it has a direct implication on investment that, what sort of investment we want to do. We need to we like to look at another quickly another aspects of the thing.


(Refer Slide Time: 22:20)



Value of Utility Pricing

- As mentioned before, economy of scale might not be very effective
- But cloud services don't need to be cheaper to be economical!
- Consider a car
 - Buy or lease for INR 500/- per day
 - Rent a car for INR 5,000/- per day
 - If you need a car for couple of days in a trip, buying would be much more costly than renting
 - **It depends on the demand**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



What we say value of utility pricing right. That how things will be priced, how things will be there, how I can provision system and it may also help me to look at that, whether it is useful to be go to cloud or having something at your own premises.

So, economy of scale might not be very effective always means all taking consideration right, but cloud service do not need to be cheaper to be economical right. So, it is economic cop thinks, may be based on that my requirement and etcetera what sort of demand I am going. Like, if we the popular example buildings to give is that a consider a car. So, buy or leasing a car may cost me something 500 per day right. Whereas, renting a car may be say for example, 500 per day. So, when it is economical? Suppose, I by looking at it is always the buying the car may be economical.



But, I if I am commuting say large distance or say one scene; that means, one once in a month or couple of delayed days in a month. Then buying of car may be more costly than renting a car, but if I require that car on a daily basis, that going to my workplace traveling a large distance etcetera then the buying a car, will maybe economical than renting a car; all right.

So, it all depends that, what sort of demand demands you are having.

(Refer Slide Time: 24:11)

Utility Pricing in Detail

D(t)	demand for resources $0 < t < T$	$C_T = \int_0^T U \times B \times D(t) dt = A \times U \times B \times T$ $B_T = P \times B \times T$ <ul style="list-style-type: none"> Because the baseline should handle peak demand When is cloud cheaper than owning? $C_T < B_T \Rightarrow A \times U \times B \times T < P \times B \times T$ $\Rightarrow U < \frac{P}{A}$ <ul style="list-style-type: none"> When utility premium is less than ratio of peak demand to Average demand
P	$\max(D(t))$: Peak Demand	
A	$\text{Avg}(D(t))$: Average Demand	
B	Baseline (owned) unit cost [B_T : Total Baseline Cost]	
C	Cloud unit cost [C_T : Total Cloud Cost]	
U (=C/B)	Utility Premium [For rental car example, $U=4.5$]	

So, you just to do some simple some mathematic expression little bit simplified so that try to have. Like suppose, I have a demand D(t). So, demand varies over time 0 to capital

T 0 to time. So, demand for resources $D(t)$, P is the max demand or peak demand, A is the average demand. So, average over the time scale, B is the baseline or own unit cost. So, if I only unit cost, what is the baseline cost C is the cloud unit cost. So, what is the cloud unit cost if I purchase the thing and U utility premium rent, a car for that there is a some type. So, there is a rent that car maybe something, it will come whatever in our case is 5000 by 500. So, something 10 is the utility.

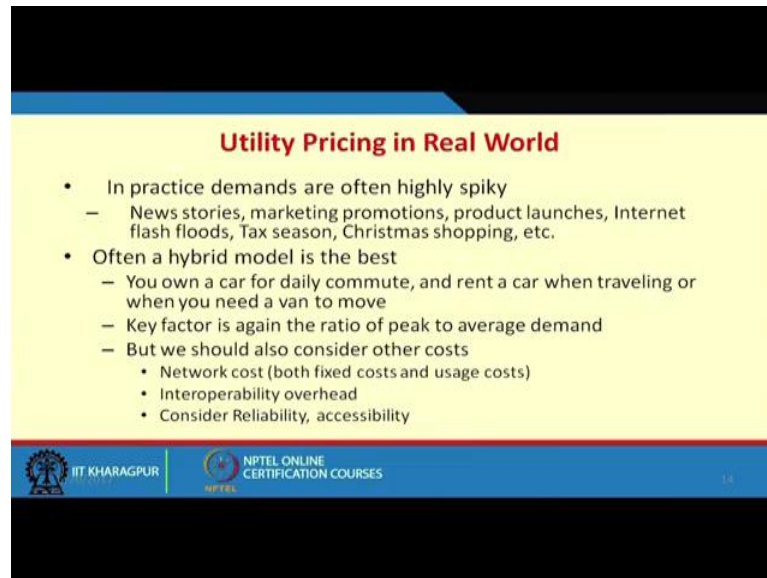
So, utility of the premium is that taking a cloud service divided by the baseline service right. That is the utility premium, we are having the things right. That is the utility premium now so, I have a variable demand $D(t)$, I have a peak demand P , we have a average demand, there is a baseline owned unit cost d_c , cloud unit cost C . And I want to find out the utility premium that is C/B . Now, if I want to see the overall cloud cost. So,

$\int_0^T U \times B \times D(t) dt$ is the overall cloud cost which is somewhat u about is the overall costing of my cloud.

If I want to calculate the overall baseline over time T ; so, P that is the peak demand into because whenever I am going for based my baseline. I have to go for the peak demand thing to calculate the thing B is the baseline own unit cost and T is the overall time scale.

Now, when cloud is cheaper, when the $C_T < B_T$, if the cost of cloud is less than the cost of baseline then it is cheaper. So, or in other sense if you look at it very simplified form. So, when P/A is greater than the utility premium right. Peak cost by average, cost a peak demand by average demand is greater than the utility premium. So, when the utility premium is less than the ratio of peak demand, to average demand then your cloud may be cheaper then owning the infrastructure or owning the services.

(Refer Slide Time: 27:05)



Utility Pricing in Real World

- In practice demands are often highly spiky
 - News stories, marketing promotions, product launches, Internet flash floods, Tax season, Christmas shopping, etc.
- Often a hybrid model is the best
 - You own a car for daily commute, and rent a car when traveling or when you need a van to move
 - Key factor is again the ratio of peak to average demand
 - But we should also consider other costs
 - Network cost (both fixed costs and usage costs)
 - Interoperability overhead
 - Consider Reliability, accessibility

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, utility pricing in real world, in praxis demands are often offense pipe spiky like new stories suddenly, it came into things. Market promotions, product launches, internet flash flood something goes to the internet etcetera. Some seasonal things like Christmas, Tax. So, those time things goes up. Often hybrid model is based. Like you own a car for daily commute, but rent a car when traveling or when you need a larger move to a larger distance to move. Key factor is again the ratio of peak to average demand. So, key factor is there again that, what is the ratio between the peak to average demand, but we should also consider other cost.

Like which had not considered in the previous calculation, then our network cost both fixed and usage costs interoperability overhead Like different one information, a one data is talking to other there in overhead. Consider reliability accessibility so and so far right. So, these are the different factors.

(Refer Slide Time: 28:17)

Value of on-Demand Services


- Simple Problem: When owning your resources, you will pay a penalty whenever your resources do not match the instantaneous demand
 - i. Either pay for unused resources, or suffer the penalty of missing service delivery

$D(t)$ – Instantaneous Demand at time t
 $R(t)$ – Resources at time t

Penalty Cost $\propto \int |D(t) - R(t)| dt$

- If demand is flat, penalty = 0
- If demand is linear periodic provisioning is acceptable

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Another aspect, we will just quickly see the value of on demand services. So, simple problem when owning your own resources, you pay the penalty whenever the resources do not match with the instantaneous demand. Suppose, I have 100 resources and I pay the penalty. When it is underutilized, suppose it is utilized with 80 percent, 70 percent then I pay the penalty for the rest of the 10 to a 20, 30 things, right. Either pay for unused resources or suffer penalty of missing service delivery things are there right. So, if it is higher things then I do not able to service.




So, penalty is how do I calculate? Penalty is proportional to $\int |D(t) - R(t)| dt$, all right. If demand is flat penalty is 0 rights. If demand is linear periodic provisioning is acceptable, right if it is a linear than periodic provisioning the acceptable.

(Refer Slide Time: 29:17)

Penalty Costs for Exponential Demand

- Penalty cost $\propto \int |D(t) - R(t)| dt$
- If demand is exponential ($D(t)=e^t$), any fixed provisioning interval (t_p) according to the current demands will fall exponentially behind
- $R(t) = e^{t-t_p}$
- $D(t) - R(t) = e^t - e^{t-t_p} = e^t(1 - e^{-t_p}) = k_1 e^t$
- Penalty cost $\propto c.k_1 e^t$

Exponential Growth with Continuous Monitoring And Non-Zero Provisioning Interval

If the demand is non-linear, then periodic provisioning in cloud is a big question right. Suppose if the demand is exponential like in this case $D(t) = e^t$ right, any fixed provisioning time interval t_p according to the current demand we will feel you will fall exponentially behind. Suppose, I require time t_p to provision the things right by the time I provision it has gone up. So, it goes on, this at t equal to e of t minus t_p . This time D_t that time to provisioning t_p , will create a havoc right. So, $D(t) - R(t) = k_1 e^t$ if you say.

So, penalty of cost is $c.k_1 e^t$. In other sense this penalty grows exponentially. It is extremely difficult to match this, unless you over provision and try to look at that is also at times difficult because it grows in exponential things. So, if you need to be careful that, when what type of demands we are expecting need to study and based on that the provisioning should be there.

(Refer Slide Time: 30:27)

Assignment 1

Consider the peak computing demand for an organization is 120 units. The demand as a function of time can be expressed as:

$$D(t) = \begin{cases} 50 \sin(t), & 0 \leq t < \pi/2 \\ 20 \sin(t), & \pi/2 \leq t < \pi \end{cases}$$

The resource provisioned by the cloud to satisfy current demand at time t is given as:

$$R(t) = D(t) + \delta \cdot \left(\frac{dD(t)}{dt}\right)$$

Where, δ is the delay in provisioning the extra computing resource on demand

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, based on these I have I kept a small assignment, for you to work at your own time and we will discuss this assignment in one of these classes during some free time. So, it says that consider a peak computing demand of an organization 120 units and demand of the functions time express. At this like these, are the demand of with respect $D(t)$ is that demand over t , the resource provisioning of the cloud to satisfied the current demand t is at equal to so and so far where δt is the delay in provisioning extra computing resource on demand. So, δ this tilde is the delay, right.

(Refer Slide Time: 31:11)

Assignment 1 (contd...)

The cost to provision unit cloud resource for unit time is 0.9 units. Calculate the penalty and draw inference.

[Assume the delay in provisioning is $\pi/12$ time units and minimum demand is 0]

(Penalty: Either pay for unused resource or missing service delivery)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what we say that, the cost to provision unit cloud resource for unit time is 0.9 units. So, we want to calculate the penalty, if any and draw the inference like why this type of penalty how things are there it is so and so far. So, we are assuming some of the factors like, what sort of the; what should be the delay in provisioning, penalty it may be either pay for on his users or missing service delivery and type of things. So, this I encourage you to look at this problem, with this thing that while things will be there will in subsequent class one of this class, we will discuss this problem. So, that is all for today. So, we looked at out the economy of cloud and where it is beneficial and what are the different aspect to which drives the economy of cloud and type of things.

Thank you.