Cloud Computing : Economics

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.

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.

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.

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.

A Useful Measure of “Smoothness”

- The coefficient of variation C_V
  - ➔ ≠ the variance σ^2 nor the correlation coefficient
- Ratio of the standard deviation σ to the absolute value of the mean |μ|
- “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

Coefficient of variation C_v

- X_1 , X_2 , ..., X_n independent random variables for demand
  - ➔ Identical standard variation σ and mean μ
- Aggregated demand
  - ➔ Mean ---> sum of means: n. μ
  - ➔ Variance ---> sum of variances: n. σ^2
  - ➔ Coefficient of variance ---> $(1 / (n)^{1/2})$. C_v
- Adding n independent demands reduces the C_v by $(1 / (n)^{1/2})$
  - ➔ Penalty of insufficient / excess resources grows smaller
  - ➔ Aggregating 100 workloads bring the penalty to 10%

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 σ^2 (X)
  - Mean: n.μ, standard deviation: n.σ(X)
  - Coefficient of Variance remains constant
- Simultaneous peaks

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 computers and components
  - ➔ Locating near cheap power supplies
  - ➔ Early entrant automation tools ---> 3rd parties take care of it

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

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 10,000/- per day
  - Rent a car for INR 45,000/- a day
  - If you need a car for 2 days in a trip, buying would be much more costly than renting
    - It depends on the demand

Utility Pricing in Detail

D(t)	Demand for resources 0<t<T
P	max (D(t)) : Peak Demand
A	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]

$$C_T = \int_0^T U \times B \times D(t)dt = A \times U \times B \times T$$

$$B_T = P \times B \times T$$

- Because the baseline should handle peak demand

When is cloud cheaper than owning?

$$C_T < B_T \rightarrow A \times U \times B \times T < P \times B \times T$$
$$\rightarrow U < \frac{P}{A}$$

- When utility premium is less than ratio of peak demand to Average demand

Utility Pricing in Real World

- In practice demands are often highly spiky
  - News stories, marketing promotions, product launches, Internet flash floods (Slashdot effect), tax season, Christmas shopping, processing a drone footage for a 1 week border skirmish, 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

Value of on-Demand Services

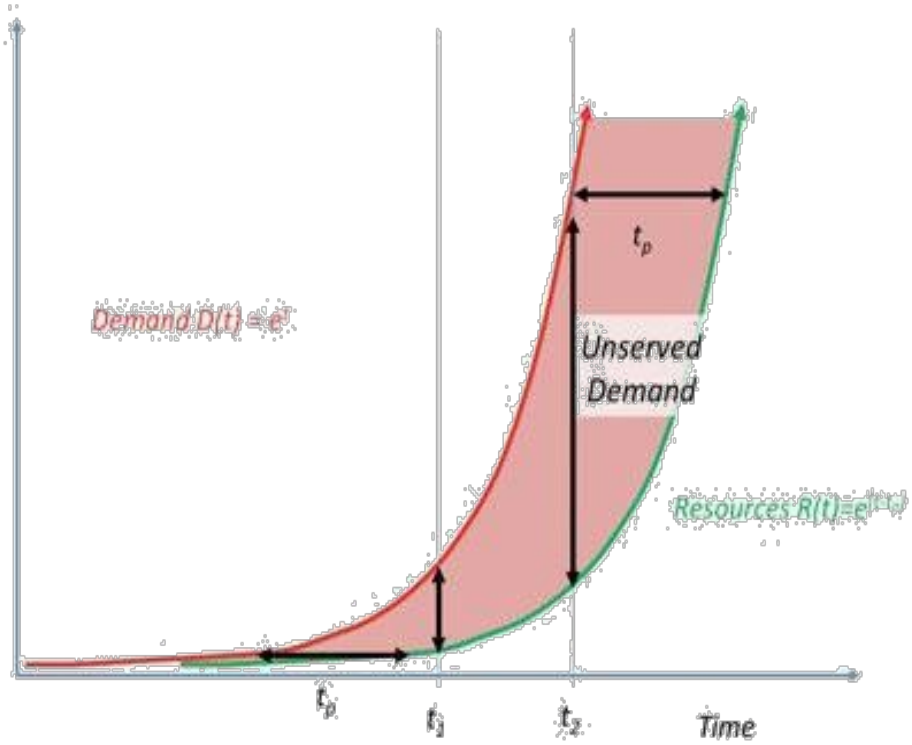
- Simple Problem: When owning your resources, you will pay a penalty whenever your resources do not match the instantaneous demand
  - Either pay for unused resources, or suffer the penalty of missing service delivery
    - D(t) – Instantaneous Demand at time t
    - R(t) – Resources at time t

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

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

Value of on-Demand Services

- 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
- $R(t) = e^t - t_p$
- $D(t) - R(t) = e^t - e^{t-t\,p} = e^t\,(1 - e^{tp}) = k_1 e^t$
- Penalty cost $\propto c.k_1\,e^t$



Exponential Growth with Continuous Monitoring And Non-Zero Provisioning Interval