Quality-Assured Cloud Bandwidth Auto-Scaling for Video-on-Demand Applications

Di Niu, Hong Xu, Baochun Li
University of Toronto

Shuqiao Zhao
UUSee, Inc., Beijing, China
Applications in the Cloud

Netflix moved to Amazon Web Services in 2010
Traditional Enterprise Operation

Server Capacity

Over-provision

Demand

Days

0 1 2
Auto Scaling in the Cloud

Reduced cost: from infrastructure investment to metered billing (pay-as-you-go)
No bandwidth guarantee!

Unappealing to delay-sensitive applications

Video-on-Demand, Gaming
Bandwidth Reservation is becoming feasible for traffic from a VM to the Internet.


Can We AutoScale Bandwidth?

A Naive Idea:
if bandwidth utilization > 90%, reserve more bandwidth.

Passive!
No guarantee!
Apps don’t know demands!
Individual reservations are costly!
Utilize the Data and Computing Power in the Cloud
From IaaS to Platform as a Service (PaaS)

VoD Apps

Payment

Performance Guarantee

Cloud Provider

Data Center

Bandwidth Reservation Request Direction

Data Center

Learning Prediction

Predictive Quantitative Multiplex

Thursday, March 29, 2012
Roadmap

1. Demand Prediction

2. Request Direction
   Resource Reservation

Cloud Provider

Tenant

Data Center

Payment

Performance Guarantee
Data Mining

- **UUSee**: a VoD provider based in China
- **200+ GB traces, 21 days**
  - reports of online users every 10 minutes
- Bandwidth demand in each **video channel**
A Typical Video Channel

Bandwidth (Mbps)

Time periods (1 period = 10 minutes)
Prediction based on Learning

- Predict expected demand
- Estimate demand variation
Seasonal ARIMA Models

Bandwidth (Mbps)

- Original data
- 10-minute-ahead prediction

Time periods (1 period = 10 minutes)

Thursday, March 29, 2012
GARCH modeling of volatility

- Prediction error
- Estimated conditional error standard deviation

Mbps

Time periods (1 period = 10 minutes)
Prediction Results

- Expected demand of each tenant $\mu_i$
- Demand standard deviation $\sigma_i$
- Demand covariance matrix $\Sigma = [\sigma_{ij}]$
Roadmap

Request Direction

Bandwidth Reservation

Tenant

Payment

Performance Guarantee

Cloud Provider

1

Demand Prediction

$\sum \mu_i$

Data Center

Data Center

Request Direction

Bandwidth Reservation

Thursday, March 29, 2012
Individual Reservation

- Tenant $i$: random demand $D_i$
- Tenant $i$ wants to reserve $R_i$ bandwidth s.t.

$$ Pr(D_i > R_i) < \epsilon $$

- If $D_i$ is Gaussian, then

$$ R_i = \mu_i + \theta(\epsilon)\sigma_i $$
Mixing anti-correlated demands saves bandwidth reservation
The bandwidth consumption of two channels is split into two data centers, each serving a part of demand from each channel, mix them and reserve within a short period such as 10 minutes. This is hard with reservation made for each channel in each data center so that the total bandwidth reservation, even within each 10-minute period, we leverage the estimated demand correlations to optimally direct workload across data centers so that the total bandwidth reservation, even within each 10-minute period, we leverage the estimated demand correlations to optimally direct workload can be reserved from data center to data center.

We use a covariance matrix $\mathcal{C}$ to find the correlation between the demands of video channels. The bandwidth $W$ of a video channel is a random variable with expected value $\mu_W$ and variance $\sigma^2_W$. Let $\mathcal{D}$ be the symmetric demand portfolio of $N$ video channels. The aggregate demand $\mathcal{D}$ is a random variable with expected value $\mu_D$ and variance $\sigma^2_D$. We define a load direction decision as a weight matrix $\mathcal{L}$, where $\mathcal{L}_i$ is the fraction of workload of channel $i$ to be directed to data center $A_i$. The optimal load direction is given by $\mathcal{L}^*$, which minimizes the total bandwidth reservation, even within each 10-minute period.

Throughout the paper, we assume that $\mathcal{D}$ is a Gaussian random variable with mean $\mu_D$ and variance $\sigma^2_D$. The coefficient of variation $\rho_D = \sigma_D / \mu_D$ is the ratio of the standard deviation to the mean. We use $\mathcal{L}^*$ to denote the optimal load direction decision.

We consider an individual time period, without loss of generality, suppose that the correlation between the demands of video channels is uniformly distributed. The bandwidth $W$ of a video channel is a random variable with expected value $\mu_W$ and variance $\sigma^2_W$. Let $\mathcal{D}$ be the symmetric demand portfolio of $N$ video channels. The aggregate demand $\mathcal{D}$ is a random variable with expected value $\mu_D$ and variance $\sigma^2_D$. The coefficient of variation $\rho_D = \sigma_D / \mu_D$ is the ratio of the standard deviation to the mean. We use $\mathcal{L}^*$ to denote the optimal load direction decision.

We use $\mathcal{L}^*$ to denote the optimal load direction decision. The optimal load direction decision decision is such that $\mathcal{L}^* = \mathcal{C}^{-1}$, where $\mathcal{C}$ is the covariance matrix of the demand portfolio. The bandwidth $W$ of a video channel is a random variable with expected value $\mu_W$ and variance $\sigma^2_W$. Let $\mathcal{D}$ be the symmetric demand portfolio of $N$ video channels. The aggregate demand $\mathcal{D}$ is a random variable with expected value $\mu_D$ and variance $\sigma^2_D$. The coefficient of variation $\rho_D = \sigma_D / \mu_D$ is the ratio of the standard deviation to the mean. We use $\mathcal{L}^*$ to denote the optimal load direction decision.

In this section, we focus on the load optimizer. Suppose $\mathcal{D}$ is a random variable with mean $\mu_D$ and variance $\sigma^2_D$. Let $\mathcal{L}^*$ be the optimal load direction decision, where $\mathcal{L}^*_i$ is the fraction of workload of channel $i$ to be directed to data center $A_i$. The optimal load direction decision decision is such that $\mathcal{L}^* = \mathcal{C}^{-1}$, where $\mathcal{C}$ is the covariance matrix of the demand portfolio. The bandwidth $W$ of a video channel is a random variable with expected value $\mu_W$ and variance $\sigma^2_W$. Let $\mathcal{D}$ be the symmetric demand portfolio of $N$ video channels. The aggregate demand $\mathcal{D}$ is a random variable with expected value $\mu_D$ and variance $\sigma^2_D$. The coefficient of variation $\rho_D = \sigma_D / \mu_D$ is the ratio of the standard deviation to the mean. We use $\mathcal{L}^*$ to denote the optimal load direction decision.

Throughout the paper, we assume that $\mathcal{D}$ is a Gaussian random variable with mean $\mu_D$ and variance $\sigma^2_D$. The coefficient of variation $\rho_D = \sigma_D / \mu_D$ is the ratio of the standard deviation to the mean. We use $\mathcal{L}^*$ to denote the optimal load direction decision.
Statistical Bin Packing (Integer Programming)

M. Wang, X. Meng, and L. Zhang,
Consolidating Virtual Machines with Dynamic Bandwidth Demand in Data Centers
INFOCOM 2011 Mini-Conference

A. Epstein, D. Breitgand
Improving Consolidation of Virtual Machines with Risk-aware Bandwidth Oversubscription in Compute Clouds
INFOCOM 2012 Mini-Conference
Problem Formulation: \[ \text{min} \sum_s A_s \]

s.t. \[ A_s \leq C_s, \]
\[ \Pr(L_s > A_s) \leq \epsilon, \]
\[ \sum_s w_{si} = 1. \]

Total bandwidth reservation
Capacity of datacenter \( s \)
Load of datacenter \( s \): \[ L_s = \sum_i w_{si} D_i \]
All \( D_i \) served

Reservation
Request Direction (Fractional in [0,1])
Random Demand

\( A_1 \)
Data Center
\( D_1 \)

\( A_2 \)
Data Center
\( D_2 \)
Bandwidth reservation minimization

\[
\min \sum_s A_s \\
\text{s.t. } A_s \leq C_s, \\
\Pr(L_s > A_s) \leq \epsilon, \\
\sum_s w_{si} = 1.
\]

For Gaussian demands

\[
\Pr(L_s > A_s) \leq \epsilon \quad \longrightarrow \quad \mathbb{E}[L_s] + \theta(\epsilon) \sqrt{\text{var}[L_s]} \leq A_s
\]

\[
\mathbb{E}[L_s] = \sum_i w_{si} \mu_i \quad \longrightarrow \quad \text{Expected demand } i
\]

\[
\text{var}[L_s] = \sum_{i,j} \sigma_{ij} w_{si} w_{sj} \quad \longrightarrow \quad \text{covariance between } i \text{ and } j
\]
Problem: each video is replicated to every data center!
Suboptimal Solutions

- **Per-DC Optimal**

\[
\begin{align*}
\min & \sum_s A_s \\
\text{s.t.} & A_s \leq C_s, \\
\Pr(L_s > A_s) & \leq \epsilon, \\
\sum_s w_{si} & = 1.
\end{align*}
\]

Solve for each \( s \) one by one

\[
\begin{align*}
\min & A_s \\
\text{s.t.} & A_s \leq C_s, \\
\Pr(L_s > A_s) & \leq \epsilon
\end{align*}
\]

- **Per-DC Limited Channels**

  - Add channel number constraint per data center:
    Integer Programming; Heuristic solutions
Trace-Driven Simulations
35 data centers; Each capacity 300 Mbps; 176 channels (aggregated); Peak demand: 7.55 Gbps; Mean demand: 5.62 Gbps; $\varepsilon = 2\%$
35 data centers; Each capacity 300 Mbps; 176 channels (aggregated);
Peak demand: 7.55 Gbps; Mean demand: 5.62 Gbps; $\varepsilon = 2\%$
Conclusions

- Use smart data and prediction for cloud workload management
- Probabilistic QoS guarantee and statistical multiplexing
- Similar to financial risk management
- Pricing this guaranteed service
Thank you!

Google “Di Niu”

http://iqua.ece.toronto.edu/~dniu/