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

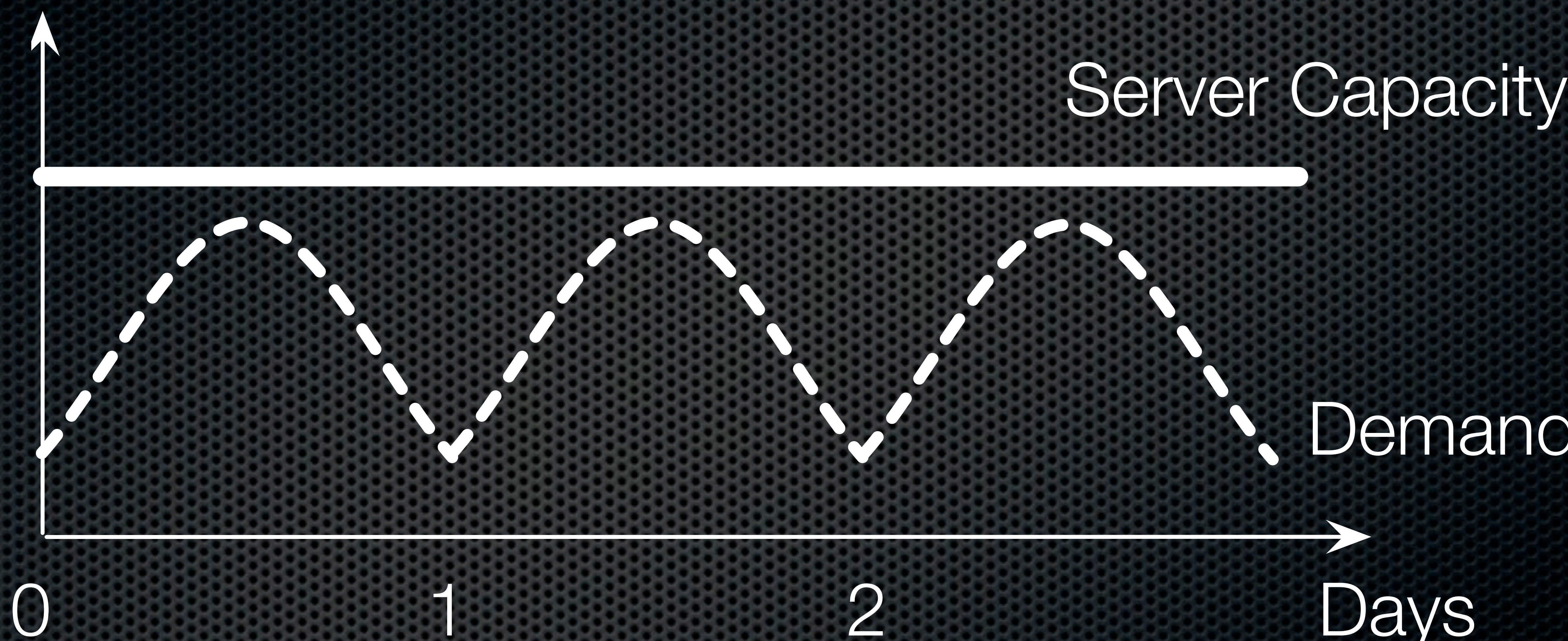
Di Niu, Hong Xu, Baochun Li
University of Toronto

Shuqiao Zhao
UUSee, Inc., Beijing, China

Applications in the Cloud

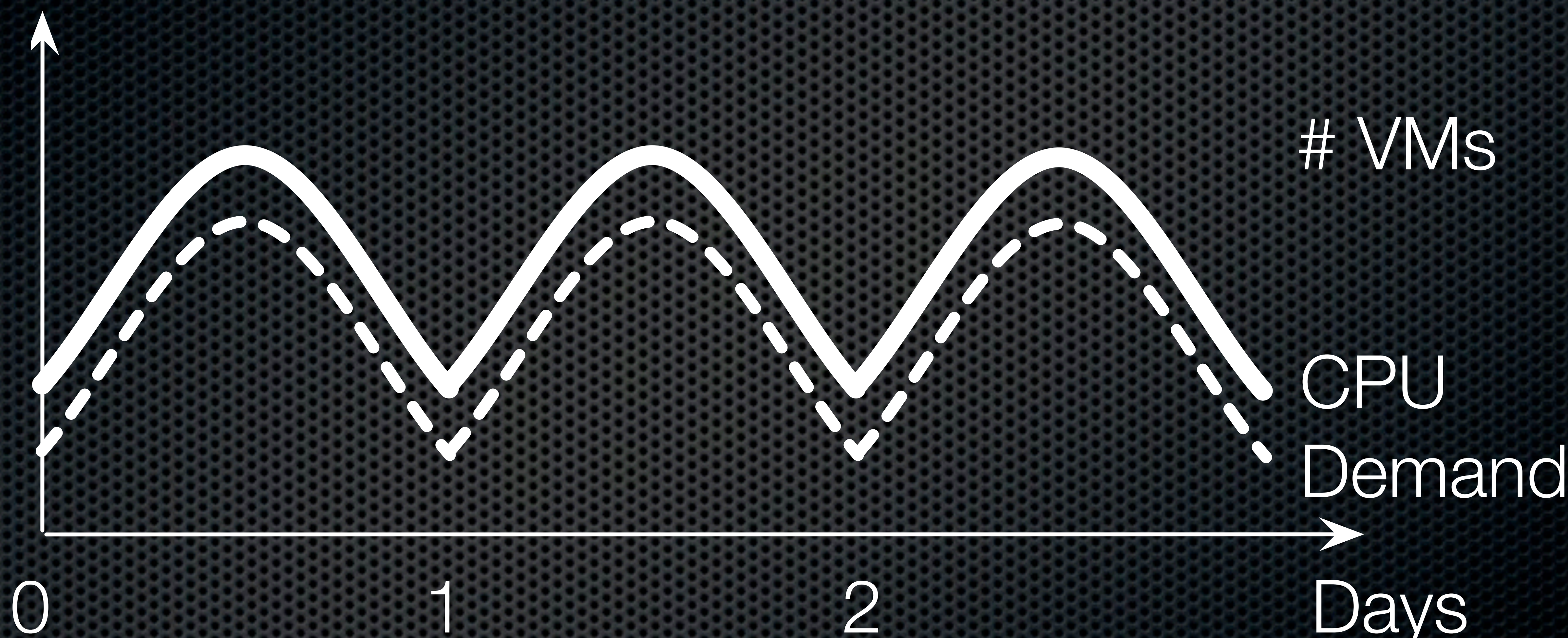
Netflix moved to Amazon Web Services in 2010

Traditional Enterprise Operation

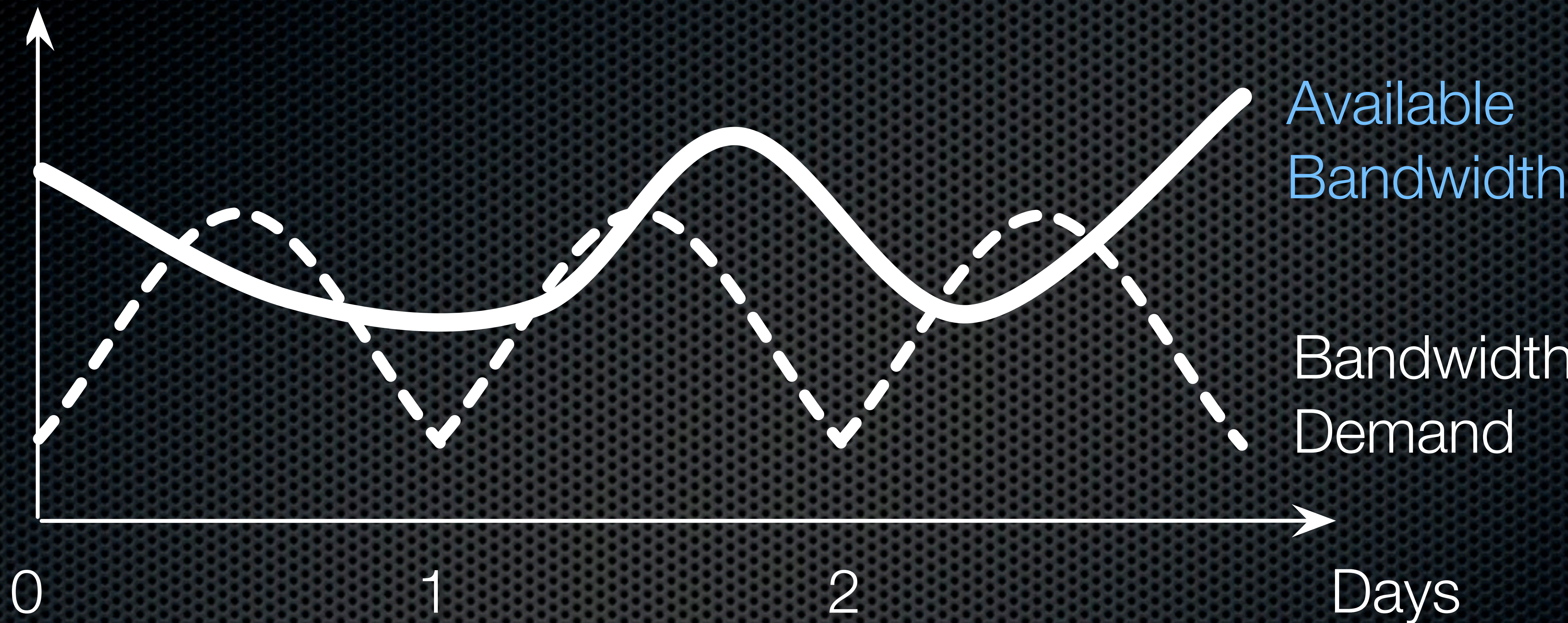


Over-provision

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

H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron,
Towards Predictable Datacenter Networks
ACM **SIGCOMM '11**

C. Guo et al.
SecondNet: a Data Center Network Virtualization Architecture with Bandwidth Guarantees
ACM **CoNEXT '10**

Can We AutoScale Bandwidth?

~~A Naive Idea:~~

~~if bandwidth utilization > 90%, reserve more bandwidth.~~

Passive!

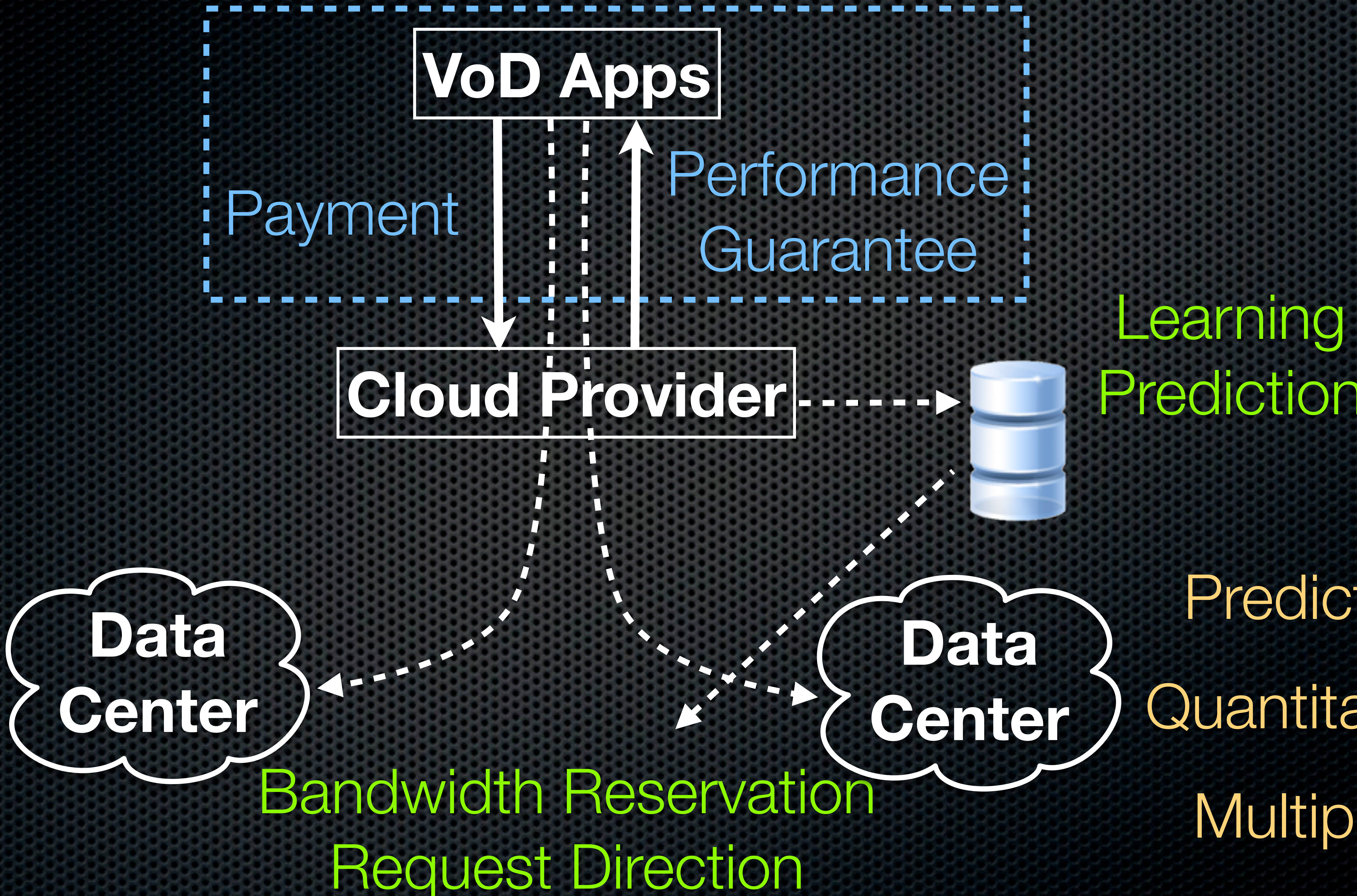
No guarantee!

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)

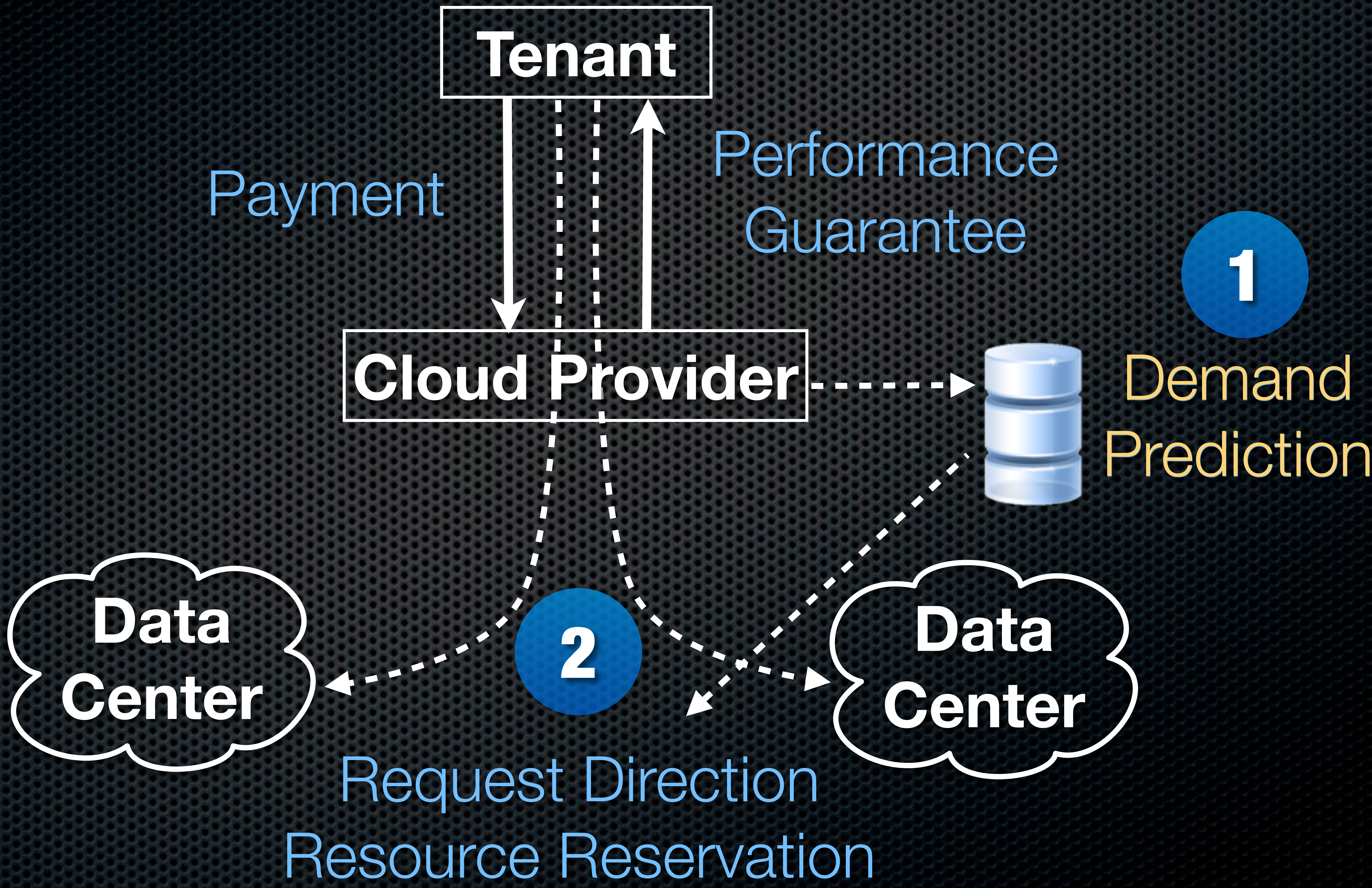


Predictive

Quantitative

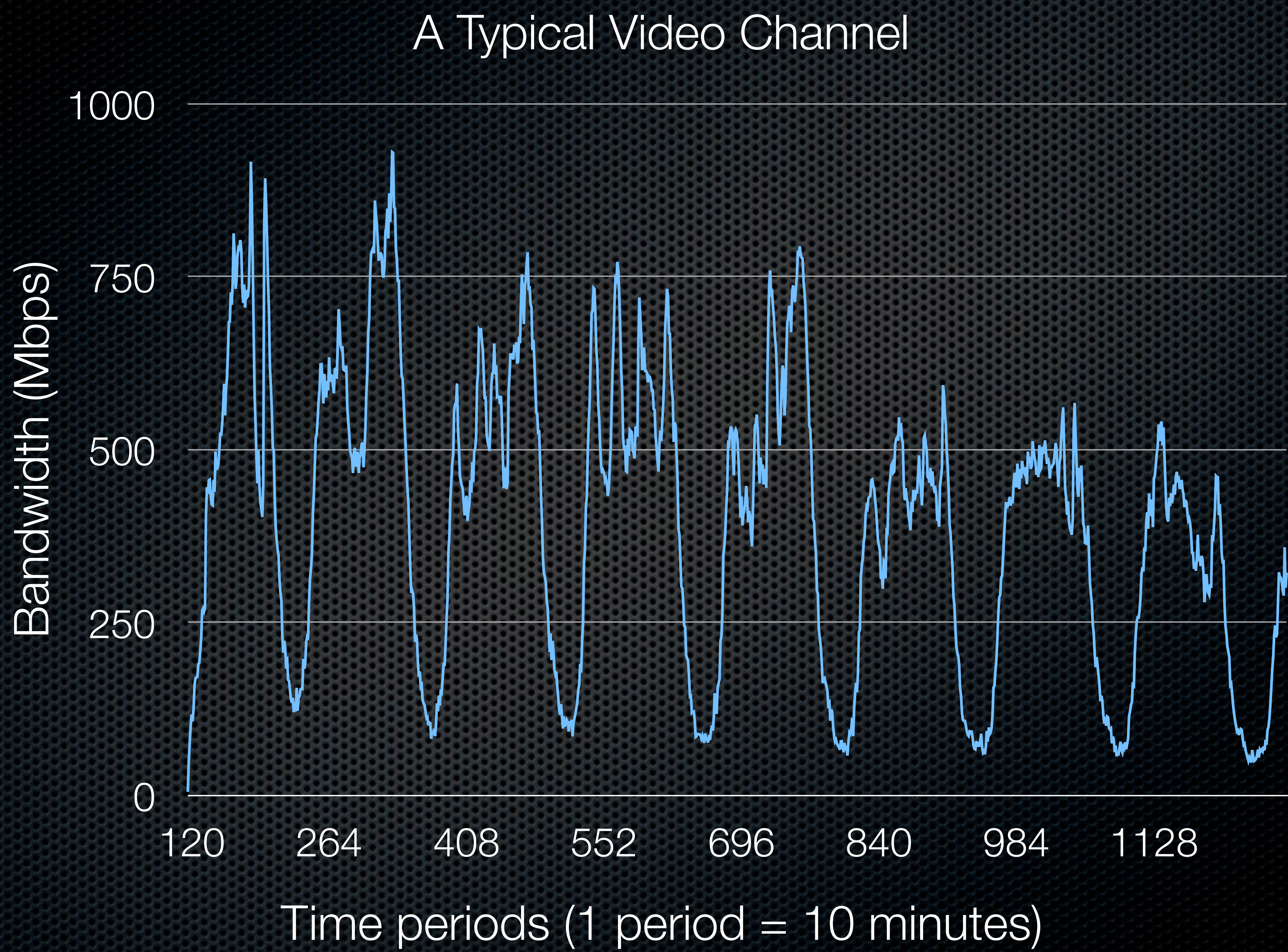
Multiplex

Roadmap


Tenant
Payment
Performance
Guarantee
1
Cloud Provider
Demand
Prediction
Data
Center
2
Data
Center
Request Direction
Resource Reservation

Data Mining

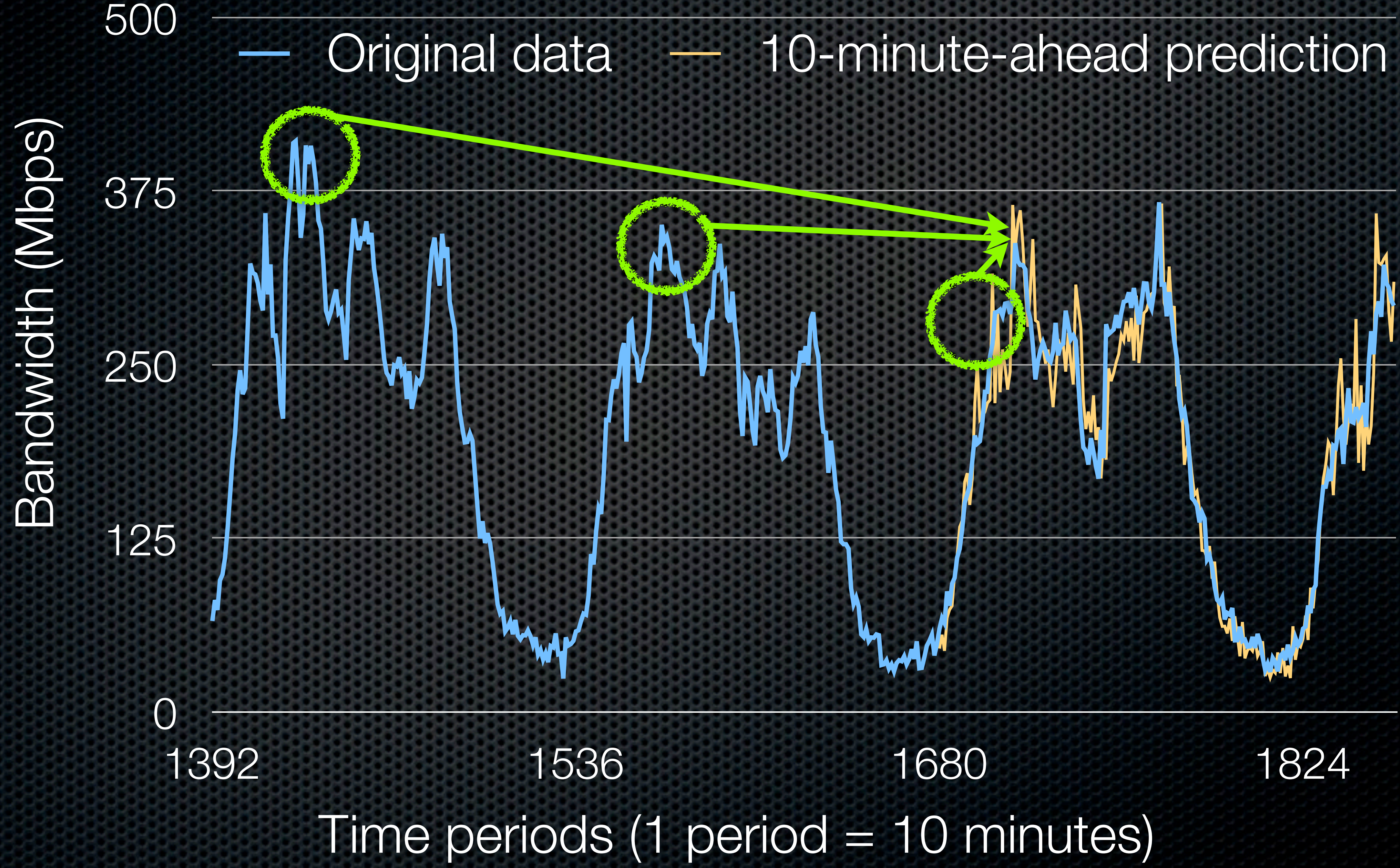
- UUSee: a VoD provider based in China
- 200+ GB traces, 21 days
  - reports of online users every 10 minutes
- Bandwidth demand in each video channel


A Typical Video Channel
1000
750
500
250
0
Bandwidth (Mbps)
120
264
408
552
696
840
984
1128
Time periods (1 period = 10 minutes)

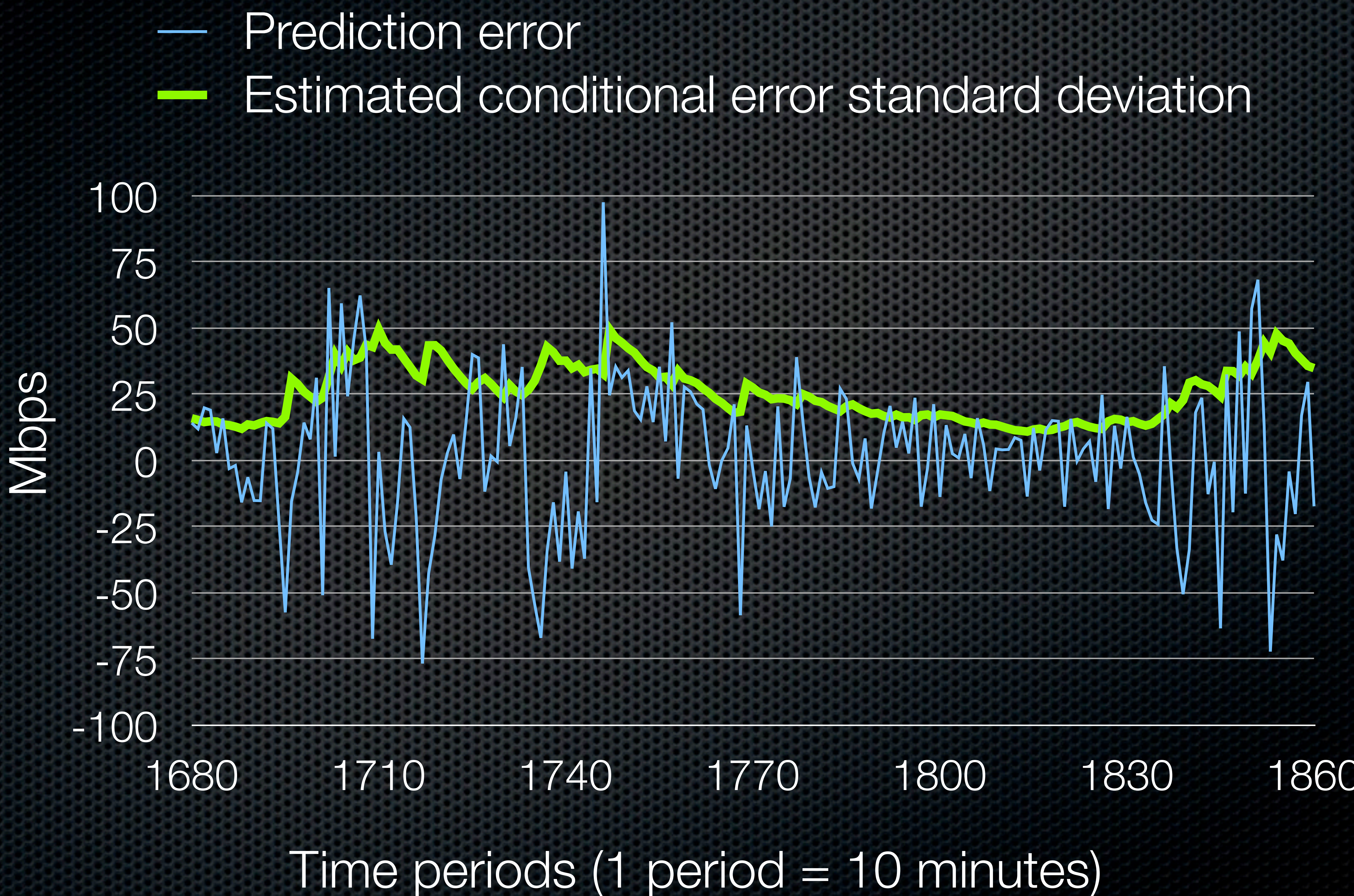
Prediction based on Learning

- Predict expected demand
- Estimate demand variation

Seasonal ARIMA Models



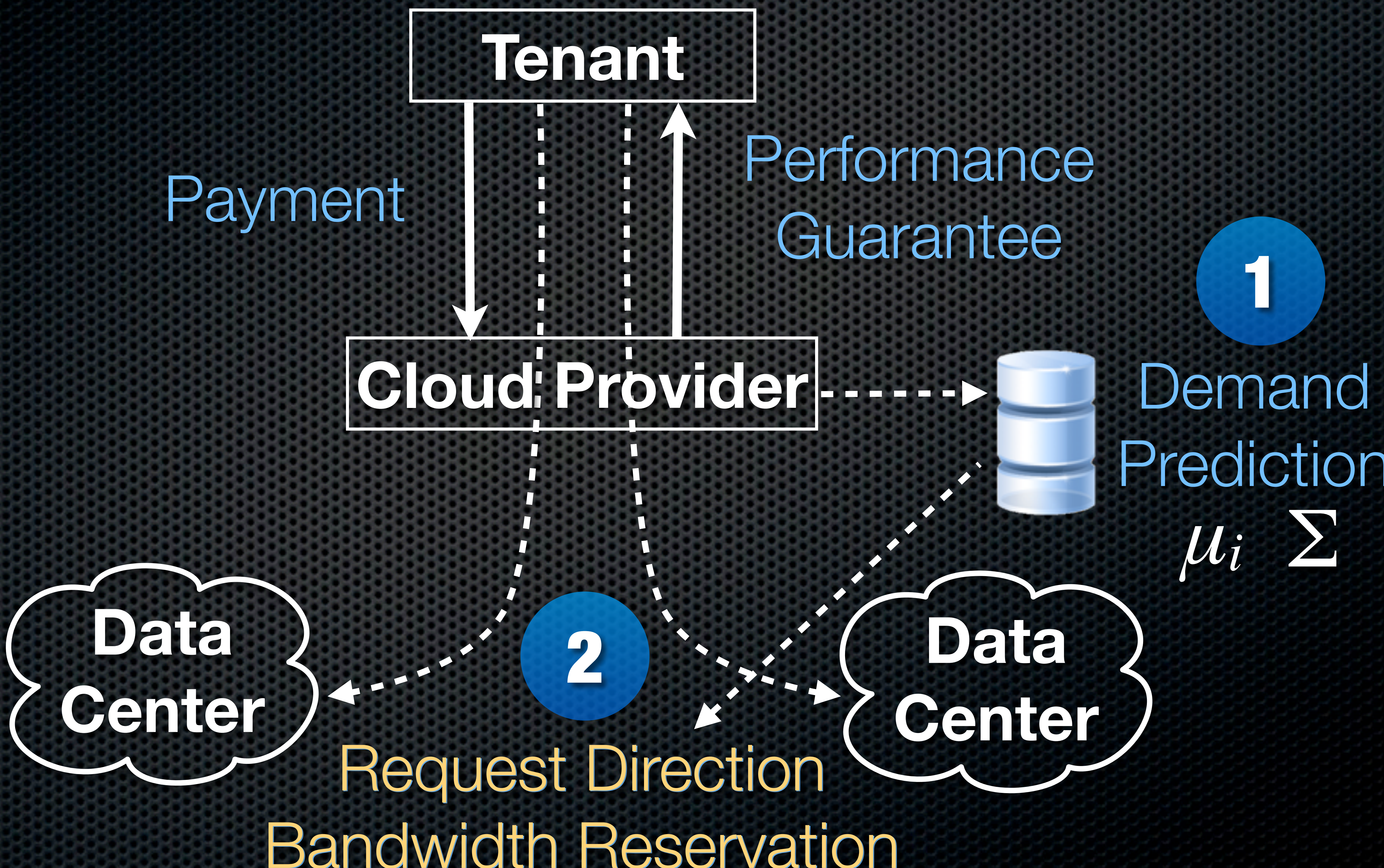
GARCH modeling of volatility



Prediction Results

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

Roadmap


Tenant
Payment
Performance Guarantee
Cloud Provider
1
Demand Prediction
μ_i Σ
2
Data Center
Data Center
Request Direction
Bandwidth Reservation

Individual Reservation

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

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

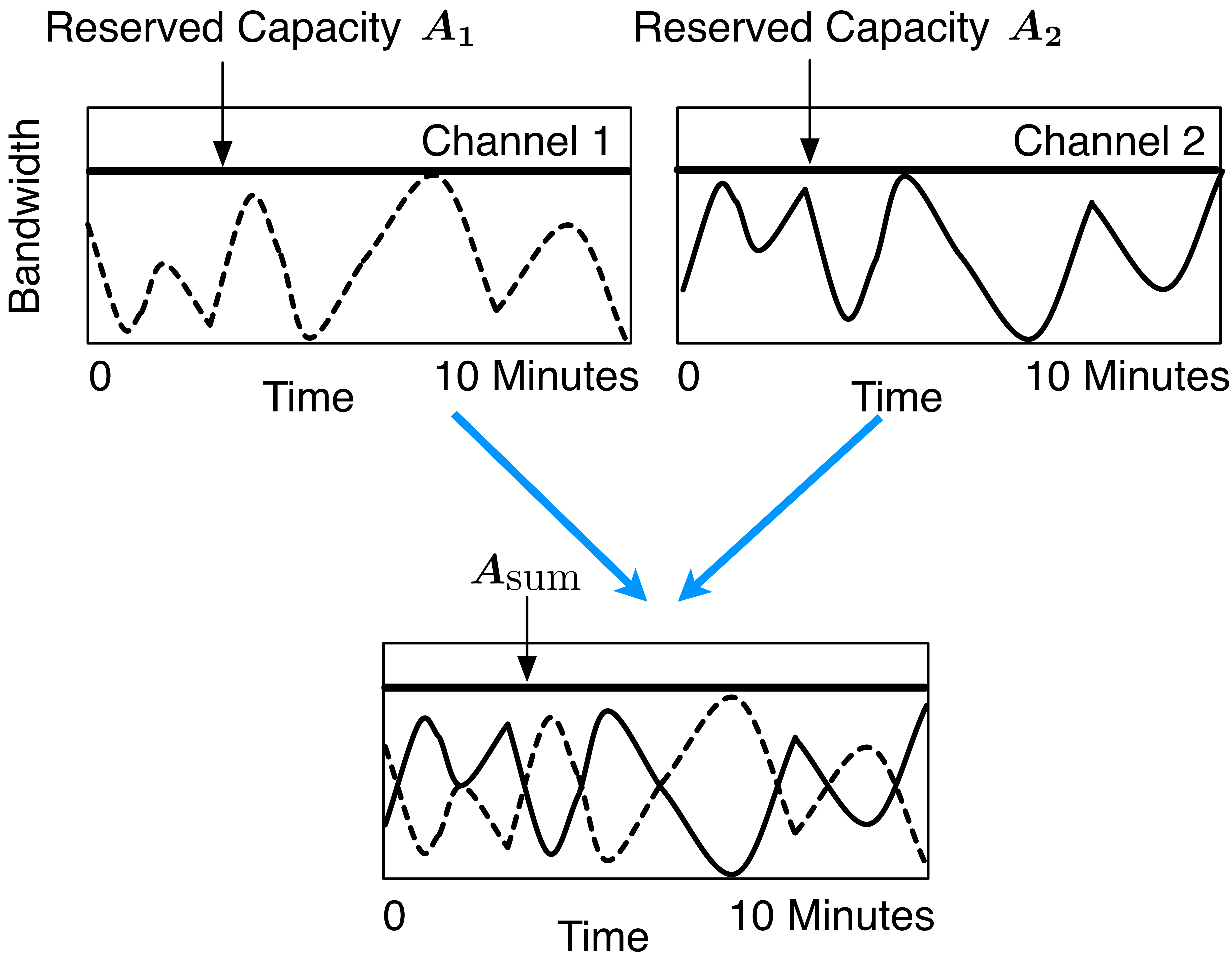
Small constant
Service Level Agreement

- If D_i is Gaussian, then

$$R_i = \mu_i + \underbrace{\theta(\epsilon)}_{\text{A constant}}\sigma_i$$

A constant

Mixing anti-correlated demands saves bandwidth reservation


Reserved Capacity A_1
Reserved Capacity A_2
Bandwidth
Channel 1
Channel 2
0
Time
10 Minutes
0
Time
10 Minutes
A_{sum}
0
Time
10 Minutes

Statistical Bin Packing (Integer Programming)

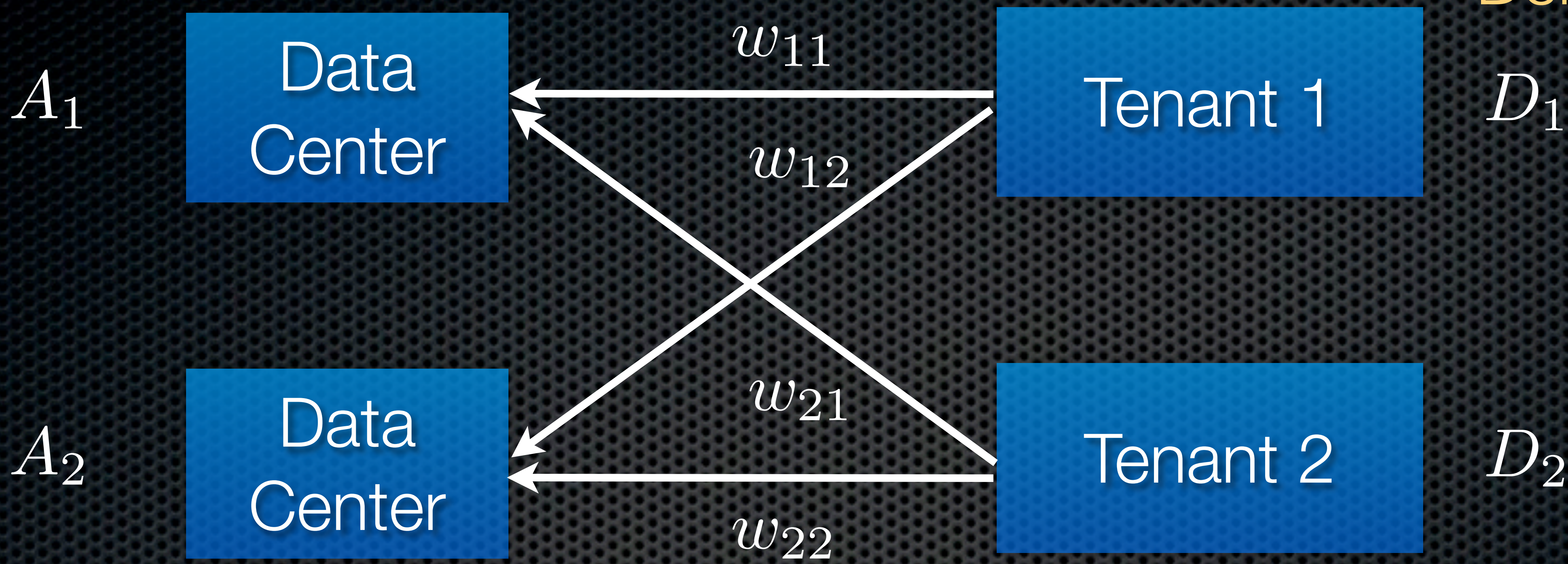
M. Wang, X. Meng, and L. Zhang,
Consolidating Virtual Machines with Dynamic Bandwidth Demand in Data Centers
INFOCOM 2011 Mini-Conference

A. Epstein, D. Breitgand
Improving Consolidation of Virtual Machines with Risk-aware Bandwidth Oversubscription in Compute Clouds
INFOCOM 2012 Mini-Conference

Reservation

Request Direction
(Fractional in $[0,1]$)

Random Demand



Problem Formulation:

$$\min \underline{\sum_s A_s}$$ → Total bandwidth reservation

$$\text{s.t.} \quad A_s \le \underline{C_s},$$ → Capacity of datacenter s

$$\Pr(\underline{L_s} > A_s) \le \epsilon,$$ → Load of datacenter s: $L_s = \sum_i w_{si} D_i$

$$\sum_s w_{si} = 1.$$ → All D_i served

Bandwidth reservation minimization

$$\min \textstyle\sum_s A_s$$
$$\text{s.t.} \quad A_s \le C_s,$$
$$\boxed{\Pr(L_s > A_s) \le \epsilon,}$$
$$\textstyle\sum_s w_{si} = 1.$$

Second order cone programming

For Gaussian demands

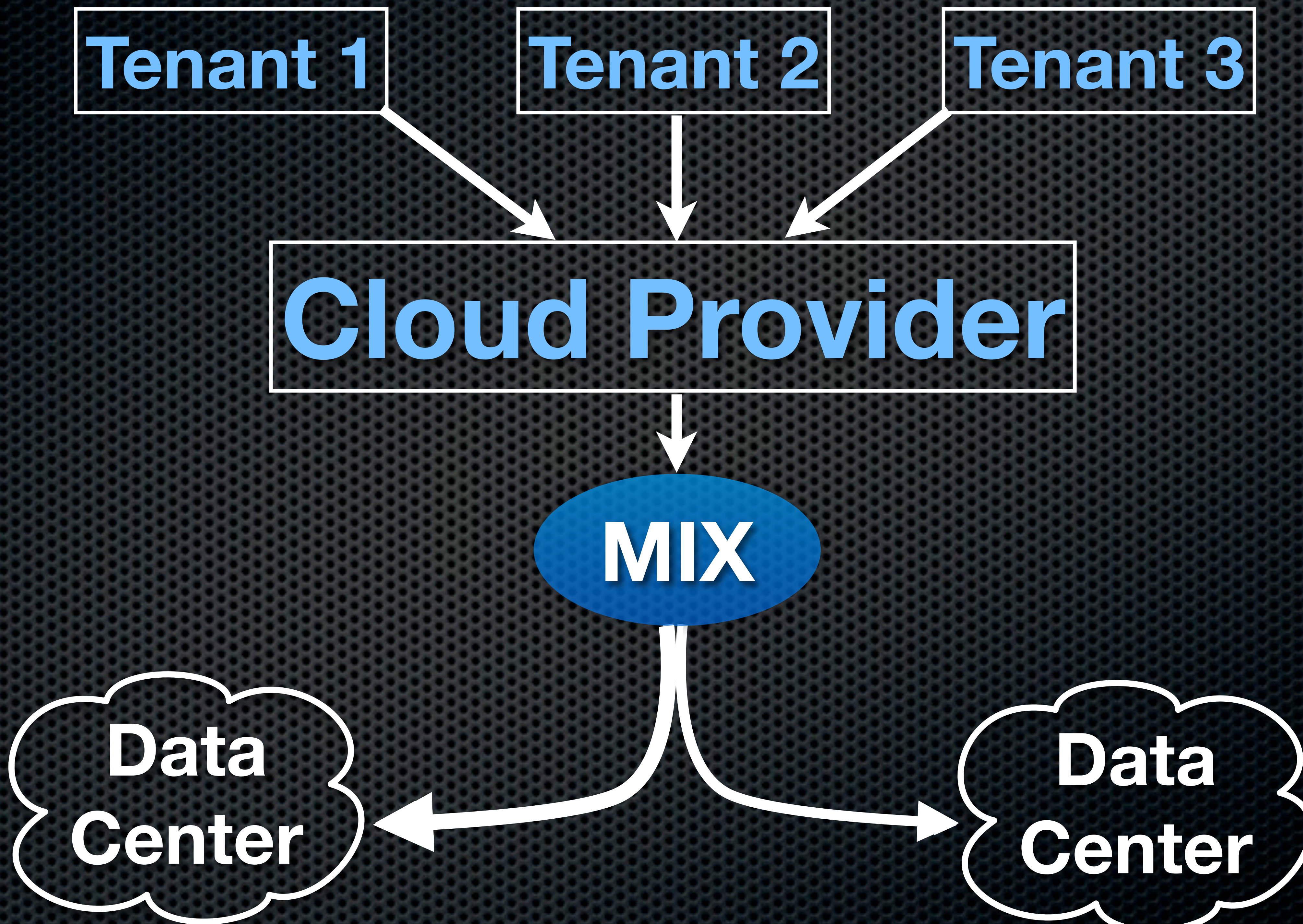
$$\Pr(L_s > A_s) \le \epsilon \longrightarrow \mathbf{E}[L_s] + \underbrace{\theta(\epsilon)}_{\text{A constant}} \sqrt{\mathbf{var}[L_s]} \le A_s$$

$$\mathbf{E}[L_s] = \sum_i w_{si} \underline{\mu_i}$$ → Expected demand *i*

$$\mathbf{var}[L_s] = \sum_{i,j} \underline{\sigma_{ij}} w_{si} w_{sj}$$

σ_{ij}: covariance between *i* and *j*

Theorem 1: Optimal Solution (Closed Form)



Problem: each video is replicated to every data center!

Suboptimal Solutions

- Per-DC Optimal

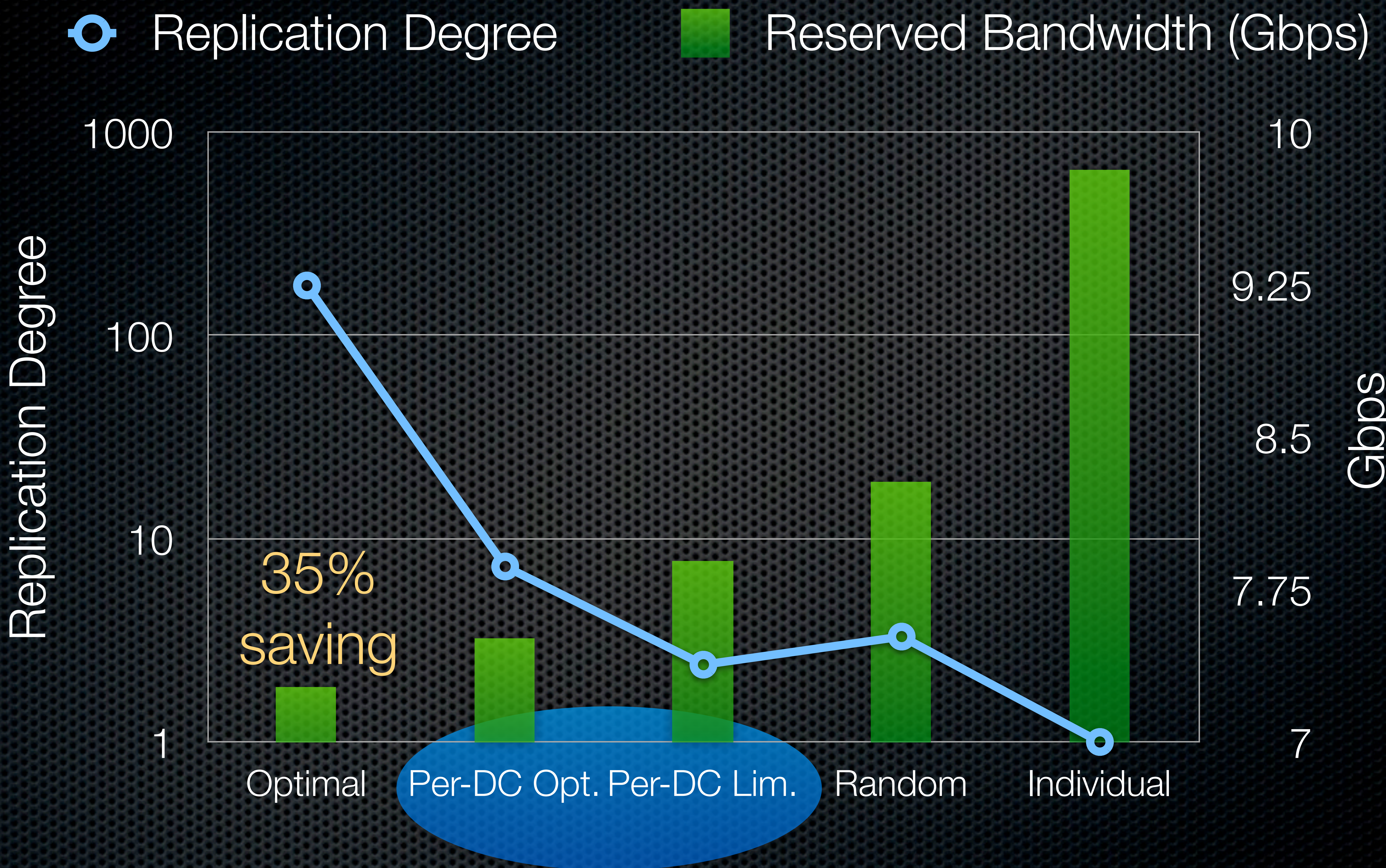
$$\begin{aligned} \min \textstyle\sum_s A_s \\ \text{s.t.} \quad A_s \le C_s, \\ \Pr(L_s > A_s) \le \epsilon, \\ \textstyle\sum_s w_{si} = 1. \end{aligned}$$

Solve for each s one by one

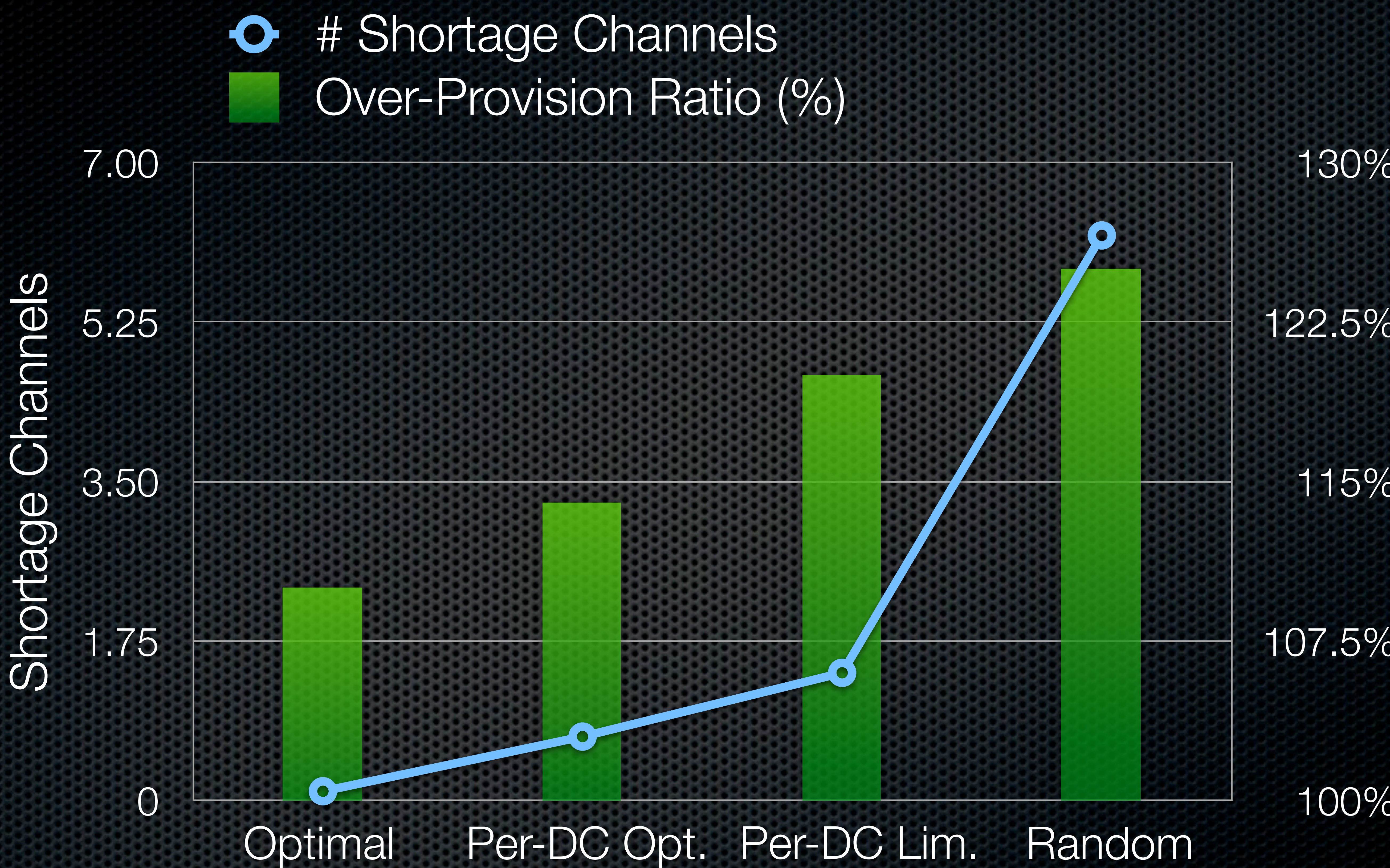
$$\begin{aligned} \min A_s \\ \text{s.t.} \quad A_s \le C_s, \\ \Pr(L_s > A_s) \le \epsilon \end{aligned}$$

- 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/