Optimal Pricing and Capacity Planning of a New Economy Cloud Computing Service Class

Jie Xu and Chenbo Zhu

Summarized by Warren Connell

3 Oct 15
Overview

- IaaS providers typically offer 2 types of services:
  - Reserved -- normal price, high uptime SLA
  - Opportunistic -- uses slack resources
    - lower price, no SLA
- High level of variance for CPU usage
  - Individual users: 30-68% of available CPU resources
  - Cluster level: 55-75% of available CPU resources
  - Slack required to guarantee high SLOs
- Is it mathematically and economically feasible to offer a middle tier?
Service Examples (Nominal)

<table>
<thead>
<tr>
<th>Type</th>
<th>SLA</th>
<th>Penalty</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reserved</td>
<td>99.95%</td>
<td>30%</td>
<td>Regular price</td>
</tr>
<tr>
<td>Opportunistic</td>
<td>None</td>
<td>None</td>
<td>Heavily discounted</td>
</tr>
</tbody>
</table>
## Service Examples (Nominal)

<table>
<thead>
<tr>
<th>Type</th>
<th>SLO</th>
<th>Penalty</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reserved</td>
<td>99.95%</td>
<td>30%</td>
<td>Regular price</td>
</tr>
<tr>
<td>Economic</td>
<td>Varies (98.9%)</td>
<td>Varies</td>
<td>Discounted</td>
</tr>
<tr>
<td>Opportunistic</td>
<td>None</td>
<td>None</td>
<td>Heavily discounted</td>
</tr>
</tbody>
</table>
Sources of Slack

- Reservation slack: capacity not yet reserved
- Allocation slack: reserved but not yet allocated
- Usage: natural fluctuations in VM operation
  - Harder to reclaim (may be used again, requires special techniques)
Questions to Answer:

• What is the optimal proportion of reclaimed resources to allocate to the new Economy class?
• What is the optimal price for the Economy class?
• Under what conditions will offering the Economy class increase the profit of the IaaS cloud service provider?
Assumptions

- Only 3 classes (Reserved/Economy/Opportunistic)
- Only 1 configuration per class
- The only SLO to worry about is availability
- There is enough demand to consume all resources
  - Best-case profit scenario
- No SLO violations for Reserved class
  - They have first priority and ability to pre-empt
  - A spike in demand causing SLO violation would affect either case equally
- Customers within a market segment are homogenous
- Prices given based on market research
Parameters / Variables

- $p_o$: the price per unit of allocated resource for the Opportunistic class;
- $A_r$: the availability SLO of the Reserved class, e.g., 99.95%;
- $A_e$: the availability SLO of the Economic class, e.g., 98.9%;
- $X$: the penalty paid to the Economy class user if the availability SLO is violated, as a discount applied to the Economy class user’s service bill;
- $V_r$: marginal valuation of the Reserved class service with the availability SLO $A_r$;
- $V_e$: marginal valuation of the Economy class service with the availability SLO $A_e$;
- $u \in [0, 1]$: the resource usage rate for both the Reserved class and the Economic class during the time period under consideration;
- $p_r$: the price per unit of allocated resource for the Reserved class;
- $p_e$: the price per unit of allocated resource for the Economic class;
- $\alpha$: the proportion of the average amount of reclaimed resources allocated to the Economy class.
Revenue Optimization
(Reserved / Opportunistic only)

\[
\max_{p_r} \pi_{ro} = p_r + p_o \int_0^1 (1 - u) f(u) \, du,
\]
subject to
\[
V_r A_r - p_r \geq 0,
\]
\[
p_r \geq p_o \geq 0.
\]
Revenue Optimization
(Reserved / Opportunistic only)

\[
\max_{p_r} \pi_{ro} = p_r + p_o \int_0^1 (1 - u) f(u) du,
\]
subject to
\[
V_r A_r - p_r \geq 0,
\]
\[
p_r \geq p_o \geq 0.
\]

\(\pi_{ro}\): total revenue
\(p_r\): price / unit (Reserved)
\(p_o\): price / unit (Opportunistic)
\(u\): resource usage rate (Reserved)
Revenue Optimization
(Reserved / Opportunistic only)

\[
\begin{align*}
\max_{p_r} & \quad \pi_{ro} = p_r + p_o \int_0^1 (1 - u) f(u) du, \\
\text{subject to} & \quad V_r A_r - p_r \geq 0, \\
& \quad p_r \geq p_o \geq 0.
\end{align*}
\]

\(V_r\): perceived value (Reserved)
\(A_r\): availability (Reserved)
\(p_r\): price / unit (Reserved)
\(p_o\): price / unit (Opportunistic)
Revenue Optimization
(with Economy class)

\[
\max_{\alpha, p_r, p_e} \pi_{reo} = p_r + p_o \int_0^{1/\alpha \mu \bar{u}} \hat{u} f(u) du \\
+ p_e \alpha \mu \bar{u} \left( 1 - X I\{\Pr[\hat{u} \geq 0] < A_e\} \right)
\]

subject to

\[
\begin{align*}
V_r A_r - p_r & \geq 0 \\
V_e A_e - p_e & \geq 0 \\
V_r A_r - p_r & \geq V_r A_e - p_e \\
V_e A_e - p_e & \geq V_e A_r - p_r \\
p_r & \geq p_e \geq p_0 \geq 0
\end{align*}
\]
Revenue Optimization
(with Economy class)

\[
\begin{align*}
\max_{\alpha, p_r, p_e} \quad & \pi_{reo} = p_r + p_o \int_0^{\frac{1}{1+\alpha\mu\hat{u}}} \hat{u} f(u) du \\
\text{subject to} \quad & V_r A_r - p_r \geq 0 \\
& V_e A_e - p_e \geq 0 \\
& V_r A_r - p_r \geq V_r A_e - p_e \\
& V_e A_e - p_e \geq V_e A_r - p_r \\
& p_r \geq p_e \geq p_o \geq 0
\end{align*}
\]

\(\pi_{reo}\): total revenue

\(p_r\): price / unit (Reserved)

\(p_o\): price / unit (Opportunistic)

\(p_e\): price / unit (Economy)
Revenue Optimization
(with Economy class)

\[
\begin{align*}
\max_{\alpha, p_r, p_e} & \quad \pi_{reo} = p_r + p_o \int_0^{1+\frac{1}{\alpha \mu \hat{u}}} f(u) \, du \\
\text{subject to} & \quad V_r A_r - p_r \geq 0 \\
& \quad V_e A_e - p_e \geq 0 \\
& \quad V_r A_r - p_r \geq V_r A_e - p_e \\
& \quad V_e A_e - p_e \geq V_e A_r - p_r \\
& \quad p_r \geq p_e \geq p_o \geq 0
\end{align*}
\]

$\alpha$: proportion of slack allocated to Economy

$\mu \hat{u}$: average slack from Reserved

$\hat{u}$: amount of reclaimed resources left over for Opportunistic

$X$: penalty for SLO violations (Economy)

$I$: indicator function

$A_e$: availability (Economy)
Revenue Optimization
(with Economy class)

\[
\begin{align*}
\max_{\alpha, p_r, p_e} & \quad \pi_{reo} = p_r + p_o \int_0^{1+\alpha \mu \mu} \hat{u} f(u) du \\
\text{subject to} & \quad V_r A_r - p_r \geq 0 \\
& \quad V_e A_e - p_e \geq 0 \\
& \quad V_r A_r - p_r \geq V_r A_e - p_e \\
& \quad V_e A_e - p_e \geq V_e A_r - p_r \\
& \quad p_r \geq p_e \geq p_o \geq 0
\end{align*}
\]

\(V_r\): perceived value (Reserved)
\(A_r\): availability (Reserved)
\(p_r\): price / unit (Reserved)
\(V_e\): perceived value (Economy)
\(A_e\): availability (Economy)
\(p_e\): price / unit (Economy)
\(p_o\): price / unit (Opportunistic)
Strategy 1
Penalty-Averse

• Try to ensure (on average) that Economy class meets SLO

• Solve: $\Pr[1 - (1 + \alpha \mu_{\bar{u}})u \geq 0] = A_e$
Strategy 1
Penalty-Averse

- Try to ensure (on average) that Economy class meets SLO
- Solve: \( \Pr[1 - (1 + \alpha \mu \bar{u})u \geq 0] = A_e \)

\[
\alpha_1^* = \frac{1}{F^{-1}(A_e) \mu \bar{u}} - \frac{1}{\mu \bar{u}}
\]
Strategy 1
Penalty-Averse

• Try to ensure (on average) that Economy class meets SLO

• Solve: \( Pr[1 - (1 + \alpha \mu_\bar{u})u \geq 0] = A_e \)

\[
\alpha_1^* = \frac{1}{F^{-1}(A_e)\mu_\bar{u}^{\alpha}} - \frac{1}{\mu_\bar{u}}
\]

\( \alpha \): proportion of slack allocated to Economy
\( \mu_\bar{u} \): average slack from Reserved
\( u \): resource usage rate (Reserved)
\( A_e \): availability (Economy)
\( \alpha_1^* \): optimal proportion
Strategy 2
Penalty-Preference

• Assume on average, that Economy class will NOT meet SLO

• Take derivatives:

\[
\frac{d\pi_{reo}}{d\alpha} = p_e \mu_\bar{u} (1 - X) - p_o \mu_\bar{u} \int_{0}^{1} \frac{1}{1 + \alpha \mu_\bar{u}} \ u f(u) \, du
\]

\[
\frac{d^2 \pi_{reo}}{d\alpha^2} = \frac{p_o \mu^2_\bar{u}}{(1 + \alpha \mu_\bar{u})^3} \left( \frac{1}{1 + \alpha \mu_\bar{u}} \right) > 0
\]

• \( \alpha^*_2 = \alpha^*_1 \) if 1st derivate negative
  • Revenue decreases as \( \alpha \) increases so use Penalty-Averse

• \( \alpha^*_2 = 1 \) if 1st derivate positive
  • Revenue increases as \( \alpha \) increases so allocate all slack
Which Strategy to Use

- Varies based on advertised SLA for Economy $A_e$ and penalty discount $X$
Which Strategy to Use

- As penalty increases, shift to penalty-averse
Which Strategy to use

- As required SLO increases, eventually shift to penalty-preference to maximize revenue
Profitability Analysis
Penalty-Averse Strategy
Profitability Analysis
Penalty-Preference Strategy
Conclusions

• Definitely room for a middle-of-the-road class between Reserved and Opportunistic service
• Actual choice/results will vary based on prices/penalties/advertised SLA
• Only a potential new pricing model
  • Actual implementation would utilize autonomous systems
Discussion