Clarifications & Errata for
“Using Burstable Instances in the Public Cloud: When and How?”

Oct. 13, 2017

This note makes some clarifications and corrects errata of [1].

Regarding the peak rate $\pi$: Solving the ODE $\pi = M + b^*w = -b + M$ gives Equ. (1) and

$$w = \frac{B^{1-z}}{T(1-z)} \text{ not } w = \frac{B^{1-z}}{T}.$$  

So, the peak rate $\pi(t) = M + (b(t))^*w$, $\pi(0) = \Pi$ and $b(0) = B$ give $\Pi = M + B^*B^{1-z}/(T(1-z))$, i.e.,

$$T = \frac{B}{(\Pi - M)(1-z)} \text{ not } T = \frac{B}{\Pi - M}.$$  

That is, the classical dual token-bucket profile indicated in Fig. 4 with a red dotted line should have a shorter duty cycle (corresponding to $B/(\Pi - M)$).

Regarding the numerically generated Fig. 16 from Equ. (6): Fig. 16 is incorrectly referred to as Fig. 5(c) just above Fig. 16, and the caption should read “until time $S = B/M$” instead of “until time $T = B/(\Pi - M)$”.

Recent empirical results show that the mean rate $\bar{\theta}$ depicted in Fig. 16 is a constant function of $s$, i.e.,

$$\forall s \leq S, \quad \bar{\theta}(s) = M.$$  

This can be algebraically shown and is consistent with a traditional mechanism for the sustainable token bucket (always at constant rate $M$), see the appendix below. That is, after idling $s$ seconds, the bucket size is $b = Ms$. Also, $\alpha = z$

for the “restored” peak rate (4). The non-constant curve in Fig. 16 is due to numerical error in measured $\alpha, \beta$.

The denominators of the fractions below Fig. 13 should be $\mu - \lambda$ not $\lambda - \mu$.

REFERENCES


This research supported in part by NSF CNS grants 1526133, 1717571, DARPA XD3 grant, and Amazon AWS Cloud Credits gift.
APPENDIX: \( \forall s \leq B/M, \bar{\theta}(s) = M \)

After \( s \) seconds idling, the token bucket occupancy \( b = Ms \). Let \( t \) be such that

\[
M + w(Ms)^2 = \pi(t) = M + (\Pi - M)(t - t/T)^\beta.
\]

This implies,

\[
Ms = B(1 - t/T)^{1/(1-z)}.
\]

Now,

\[
\bar{\theta}(s) = \frac{1}{s + T - t} \int_T^t \pi(t) \, dt
= \frac{1}{s + T - t} \left( M(T - s) + \frac{\Pi - M)T}{\beta + 1} (1 - t/T)^{\beta+1} \right)
= \frac{1}{s + T - t} \left( M(T - s) + Ms \right)
= M,
\]

where we have used the above expressions for \( w, T, sM \), and \((\beta + 1)(1 - z) = 1\).

This is consistent with simple “conservation of tokens” for a traditional token-bucket with token-rate \( M \) (and size \( B \)): At the end of every burst interval (of length \( T - t \)), the token bucket is empty. Thus, the tokens arrived in during any period (of length \( s + T - t \), \( M(s + T - t) \), must equal the amount consumed (network transmission). So, \( \bar{\theta}(s) = M(s + T - t)/(s + T - t) = M \).