MaaS: Dynamic Reliability Methods in a Clustered-Task Server

Rachel Traylor
www.themathcitadel.com
@Mathpocalypse

August 15, 2017
Mathematician, currently at Dell EMC
Research Interests: Probability Theory, Queuing Theory, Reliability Theory, Stochastic Analysis
Hobbies: Swimming away from my problems and stockpiling bacon
Why should you care?

Simplicity is a prerequisite for reliability. – Edsger Dijkstra

- server downtime is costly and makes people angry
- traffic is not steady, and workload is not constant
Goals

- Walk through an overview of mathematical solutions for server reliability
- Leave the cloud and discuss how this is useful to you
A Mathematician’s View of a Server

\[ G(w) \]

\[ \lambda \]

Server

\[ \text{Happy Completed Jobs} \]
Let’s Talk Reliability

Let $Y$ be the random variable representing the lifetime of a server.

- **Survival Function**: $S_Y(t) := P(Y > t)$
- **Hazard Function** - instantaneous rate of failure:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq Y \leq t + dt)}{dt}$$
Throwing in some randomness

\[ B(t) = r_0(t) + \sum_{j=1}^{N(t)} \mathcal{H}_j \mathbb{1}(T_j < t \leq T_j + W_j), \quad t \geq 0 \]
Survival Function for a Single Server

- random arrival times - even the arrival rate can depend on time (non-homogenous Poisson process)
- random service times with any service time distribution you want
- random stress with any distribution you want, continuous or discrete

\[
S_Y(t) = \exp \left( - \int_0^t r_0(s) ds \right) \exp \left( -E_h \left[ H \int_0^t e^{-hw} m(t-w) \tilde{G}_W(w) dw \right] \right)
\]
Changing up the model- Clustered Task Server

\[ \lambda(t) \]

\[ \begin{array}{cccc}
1 & 2 & 3 & 4 \\
G_1(t) & G_2(t) & G_3(t) & G_4(t) \\
\end{array} \]

- Customer 1
- Customer 2
- Customer 3
So many options

- independent channel selection
- dependent channel selection - constant, temporal, and random
Independent Channel Selection

- $p = p_1 = p_2 = ... = p_K$ - channel selection probability
- $G_1(w), ... G_K(w)$ - service time distributions.
- $G_W(w) = \max_i G_i(w)$
- $\eta N = \text{stress to the server, } N$ - random number of channels selected
- $\eta N \sim \text{Binomial}(K, p)$

$$S_Y(t) = \bar{F}_0(t) \exp \left(-K\eta \left[ e^{-\eta t}(1 - p + pe^{-\eta t})^{K-1} - p(1 - p)^{K-1} \right] \right) \times \int_0^t m(t - w)\tilde{G}_W(w)dw$$
Dependent Channel Selection

- Whether or not the first channel was selected affects the probability of all the other being selected
- Called *first-kind dependency*. (Other kinds detailed on The Math Citadel)
- Introduce a *dependency coefficient* $\delta \in [0, 1]$
- Can still give the survival function in closed form
Now let’s relax even $\delta$

$$\delta(t) = \frac{1}{2}\sin(x) + \frac{1}{2}$$
Now let’s relax even $\delta$

- $\lambda = \eta = 1$, $r_0 \equiv 0.01$, $p = 3/4$, $G_W(w) = e^{-w}$
- $\delta$ is a 3 state Markov Chain.
- $\delta \in \{1/10, 1/2, 9/10\}$
- Transition Matrix
  $$P = \begin{bmatrix}
  1/3 & 1/3 & 1/3 \\
  0 & 2/3 & 1/3 \\
  1/3 & 0 & 2/3
  \end{bmatrix}$$
Why do I even care about dependence? Isn’t assuming independence enough?

Figure: Percent Difference in Independent and Dependent Channel Survival Functions
Fantastic, now what?

- Let’s do something with them and talk more.
- What kind of applications do you see in this?
My questions for you to think about

▶ How would you define "stress" to a server?
▶ What metric do you use to measure traffic?
▶ Do you collect this data?
▶ Are you interested in the notion of network reliability? That is, a network built from servers that behave any of these ways?

www.themathcitadel.com