Optimal Service Elasticity in Large-Scale Distributed Systems

Noah Johnson

ELE 549 Final Project

7 December 2018

Outline

Problem Statement

Question: Given a system consisting of N queues for identical servers, and a single dispatcher, can we develop an automatic load-balancing scheme that:

a) does not rely on a centralized system queue or any global queue length information,
b) maintains constant overhead as $N \to \infty$,
c) remains competitive with related schemes

Solution: Token-Based Auto Balance Scaling (TABS)

Control Variables

Each of N servers in the system should not always be on. They

Token-Based Auto Balance Scaling (TABS)

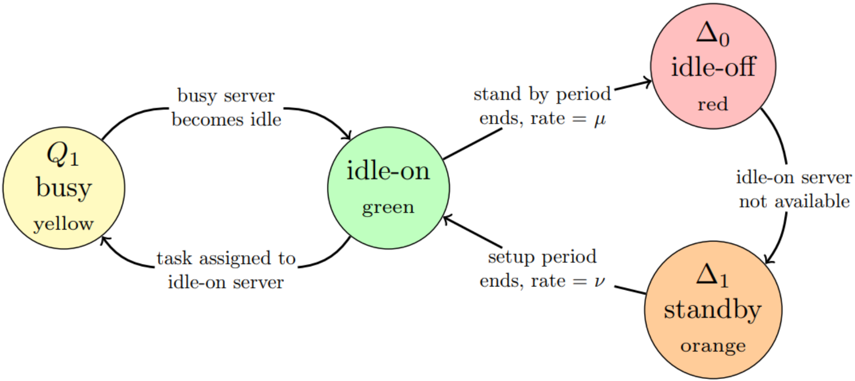
TABS is a token-based feedback protocol that tracks the state of all servers in the system

Servers can be in one of four states:

- Idle-off (server is off)
- Standby (server is in setup period)
- Idle-on (server is on, queue is empty)
- Busy (server is on, queue contains at least one job)

Protocol can be visualized by the following state diagram

TABS Decision Rules



Results

For some constant arrival rate λ and any $\mu > 0, \nu > 0$, as $N \to \infty$,

(a) $\mathbb{E}[W^{(N)}] \to 0$

(b) $\mathbb{E}[Z^{(N)}] \to 0$

By using the TABS schema, both the mean waiting time and wasted energy vanish in the limit!

Derivation of System State Space

Let:

$q_i^{(N)}(t)$: the fraction of servers with at least queue length i at time t

$\Delta_0^{(N)}(t)$: the fraction of servers in idle-off mode at time t

$\Delta_1^{(N)}(t)$: the fraction of servers in standby mode at time t

for notational simplicity, let:

$q^{(N)}(t) \triangleq (q_i^{(N)}(t)|i = 1, 2, \ldots, B)$, and $\delta^{(N)}(t) \triangleq (\delta_0^{(N)}(t), \delta_1^{(N)}(t))$

Now define E to be the space of all possible states the system can occupy:

$$E = \{(q, \delta) \in [0, 1]^{B+2}\}$$

such that:

$$q_i \geq q_{i+1} \forall i, \text{ and } \delta_0 + \delta_1 + \sum_{i=1}^{B} q_i \leq 1$$

ex. 1: System State Space for N = 2, B = 1

Q: What are the possible job states in the system at a given time t?

A: {0 jobs, 1 job, 2 jobs}

$$q^{(N=2)}(t) = \left\{\tfrac{0}{2}, \tfrac{1}{2}, \tfrac{2}{2}\right\}$$

Q: What are the possible server states in the system?

A: {both servers on, 1 in standby, 1 off}

$$\delta_0^{(N=2)}(t) = \delta_1^{(N=2)}(t) = \left\{\tfrac{0}{2}, \tfrac{1}{2}\right\}$$
$$\delta^{(N=2)}(t) = \left\{(0,0), (0, \tfrac{1}{2}), (\tfrac{1}{2}, 0), (\tfrac{1}{2}, \tfrac{1}{2})\right\}$$

Q: What is E for this scenario?

A: E is all combinations of $q^{(N=2)}(t)$ and $\delta^{(N=2)}(t)$

Deterministic Nature of System as $N \to \infty$

Assume that: $(q^N(0), \delta^N(0))$ converges to $(q^\infty, \delta^\infty)$

Proposition 3.3 : Interchange of Limits

Use `tabular` for basic tables — see Table 1, for example.

You can upload a figure (JPEG, PNG or PDF) using the files menu.

To include it in your document, use the `includegraphics` command (see the comment below in the source code).

Item	Quantity
Widgets	42
Gadgets	13

Table 1: An example table.

Readable Mathematics

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.