Asynchronous Load Balancing and Auto-scaling: Mean-field Limit and Optimal Design

Jonatha ANSELMI, Inria

13ème Atelier en Évaluation des Performances - IRIT - Toulouse
December 3, 2024

[Based on: J. Anselmi “Asynchronous Load Balancing and Auto-scaling: Mean-field Limit and Optimal Design”, IEEE/ACM Transactions on Networking, 2024]
Load Balancing and Auto-scaling

**Load balancing:**
Dispatch jobs to ON servers

**Auto-scaling:**
Scale up/down the net service capacity in response to the current load

**Challenge:** Design algorithms that achieve low wait and energy consumption
Some Examples

Supermarket checkout lines

Call centers

Data centers

In France, 10% of the electricity produced is consumed only to meet the needs of data centres [source: https://corporate.ovhcloud.com]
Serverless Computing

In the queueing literature, load balancing and auto-scaling have been mostly studied independently of each other (timescale separation assumption).
Serverless Computing

In the queueing literature, load balancing and auto-scaling have been mostly studied independently of each other (timescale separation assumption)

In serverless computing:

- a server is a software function that
  - can be flexibly instantiated in milliseconds (a time window that is comparable with the magnitude of job inter-arrival and service times)
  - No timescale separation

- Autoscaling mechanisms are extremely reactive and the decision of turning servers on are based on instantaneous observations of the current system state rather than on the long-run equilibrium behavior.
**Serverless Computing: Architectures**

Existing architectures: centralized or decentralized / synchronous or asynchronous

- **Synchronous**: Scale-up decisions taken at job arrival times (coldstarts)
- **Asynchronous**: Scale-up decisions taken independently of the arrival process

Scale-down rule: turn a server off if that server remains idle for a certain amount of time
Serverless Computing: Platforms

Centralized

- AWS Lambda, Azure Functions, IBM Cloud Functions, Apache OpenWhisk
  - Several research works

Decentralized

- Knative (Google Cloud Run)
  - Several research works

Asynchronous

- Several research works

Synchronous

- Several research works

Decentralized

- Several research works
Asynchronous Load Balancing and Auto-scaling

**Challenge 1**
To build a model to evaluate the performance of Knative
- User-defined scale-up rules
- Power-of-$d$ and JoinBelowThreshold-$d$ (JBT-$d$)

**Challenge 2**
Asymptotic Delay and Relative Energy Optimality (DREO), ie,
- the user-perceived waiting time and the relative energy wastage induced by idle servers vanish as $N \to \infty$
Markov Model
Microscopic description

Just one server:

\( f(x) \) and \( g(x) \) are the load-balancing and auto-scaling rules.

\( \lambda N \) is the job arrival rate.

\( aN \) is the rate of the auto-scaling clock.

\( \beta \) and \( \gamma \) are the server initialization and expiration rates.
Markov Model
Microscopic description

Just one server:

\( f(x) \) and \( g(x) \) are the load-balancing and auto-scaling rules

\( \lambda N \) is the job arrival rate

\( aN \) is the rate of the auto-scaling clock

\( \beta \) and \( \gamma \) are the server initialization and expiration rates

\( f(x) = \frac{1}{(Nx_{ON})} \), random dispatching

\( g(x) = \text{constant} \)

\( \beta = \infty \)

\Rightarrow \text{challenging stability region!}
Markov Model
Macroscopic description

Letting the proportion of servers with $i$ jobs and in state $j$ denoted by

$$X_{i,j}^N(t), \quad i \geq 0, \ j \in \{\text{OFF}(0), \text{INIT}(1), \text{ON}(2)\}$$

The Markov chain of interest has rates

- $x \mapsto x' := x + \frac{1}{N}(e_{i,2} - e_{i-1,2})$ with rate $\lambda N f_{i-1}(x)$
- $x \mapsto x' := x + \frac{1}{N}(e_{i-1,2} - e_{i,2})$ with rate $x_i N$
- $x \mapsto x' := x + \frac{1}{N}(-e_{0,0}, e_{0,1})$ with rate $\alpha N g$
- $x \mapsto x' := x + \frac{1}{N}(-e_{0,1}, e_{0,2})$ with rate $\beta x_{0,1} N$
- $x \mapsto x' := x + \frac{1}{N}(e_{0,0} - e_{0,2})$ with rate $\gamma x_{0,2} N$

Power-of-$d$: $f_i(x) = \frac{y_i^d - y_{i+1}^d}{y_0^d}$, \hspace{1cm} JBT-$d$: $f_i(x) = \frac{x_{i,2}}{y_0} \mathbb{I}\{\sum_{k=0}^{d} x_{k,2}=0\} + \frac{x_{i,2}}{\sum_{k=0}^{d} x_{k,2}} \mathbb{I}\{i \leq d\}$
Fluid Model and Connection with the Markov Model

**Definition 1.** A continuous function $x(t) : \mathbb{R}_+ \rightarrow S$ is said to be a fluid model (or fluid solution) if for almost all $t \in [0, \infty)$

\[
\begin{align*}
\dot{x}_{0,0} &= \gamma x_{0,2} - \alpha g \mathbb{I}_{\{x_{0,0} > 0\}} - \gamma x_{0,2} \mathbb{I}_{\{x_{0,0} = 0, \gamma x_{0,2} \leq \alpha g\}} \\
\dot{x}_{0,1} &= \alpha g \mathbb{I}_{\{x_{0,0} > 0\}} - \beta x_{0,1} + \gamma x_{0,2} \mathbb{I}_{\{x_{0,0} = 0, \gamma x_{0,2} \leq \alpha g\}} \\
\dot{x}_{0,2} &= x_{1,2} - h_0(x) + \beta x_{0,1} - \gamma x_{0,2} \\
\dot{x}_{i,2} &= x_{i+1,2} \mathbb{I}_{\{i < B\}} - x_{i,2} + h_{i-1}(x) - h_i(x),
\end{align*}
\]

(4a) (4b) (4c) (4d)

where $g := g(x) : S \rightarrow [0, 1]$, and $h_i(x) = \min\{\beta x_{0,1}, \lambda\}$ if $y_0 > 0$ and otherwise ($y_0 = 0$):

\[h_i(x) = \lambda \frac{y_i^d - y_{i+1}^d}{y_0^d}\]

(5)

if Power-of-$d$ is applied and

\[h_i(x) = \begin{cases} \\
\lambda \frac{x_{i,2}}{\sum_{k=0}^d x_{k,2}} \mathbb{I}_{\{i \leq d\}}, & \text{if } \sum_{k=0}^d x_{k,2} > 0 \\
(\beta x_{0,1} + x_{d+1,2} \mathbb{I}_{\{i = d\}}) \mathbb{I}_{\{x_{d+1,2} + (d+1)\beta x_{0,1} \leq \lambda\}}, & \text{if } \sum_{k=0}^d x_{k,2} = 0, \ i \leq d \\
\frac{x_{i,2}}{y_0} (\lambda - x_{d+1,2} - (d + 1)\beta x_{0,1})^+, & \text{if } \sum_{k=0}^d x_{k,2} = 0, \ i > d
\end{cases}\]

(6)

if JBT-$d$ is applied.
Fluid Model and Connection with the Markov Model

**Definition 1.** A continuous function \( x(t) : \mathbb{R}_+ \to S \) is said to be a fluid model (or fluid solution) if for almost all \( t \in [0, \infty) \)

\[
\begin{align*}
\dot{x}_{0,0} &= \gamma x_{0,2} - \alpha g \mathbb{I}_{\{x_{0,0} > 0\}} - \gamma x_{0,2} \mathbb{I}_{\{x_{0,0} = 0, \gamma x_{0,2} \leq \alpha g\}} & (4a) \\
\dot{x}_{0,1} &= \alpha g \mathbb{I}_{\{x_{0,0} > 0\}} - \beta x_{0,1} + \gamma x_{0,2} \mathbb{I}_{\{x_{0,0} = 0, \gamma x_{0,2} \leq \alpha g\}} & (4b) \\
\dot{x}_{0,2} &= x_{1,2} - h_0(x) + \beta x_{0,1} - \gamma x_{0,2} & (4c) \\
\dot{x}_{i,2} &= x_{i+1,2} \mathbb{I}_{\{i < B\}} - x_{i,2} + h_{i-1}(x) - h_i(x), & (4d)
\end{align*}
\]

where \( g := g(x) : S \to [0, 1] \), and \( h_i(x) = \min\{\beta x_{0,1}, \lambda\} \) if \( y_0 > 0 \) and otherwise \( (y_0 = 0) \):

\[
h_i(x) = \lambda \frac{y_i^d - y_{i+1}^d}{y_0^d} \quad (5)
\]

if Power-of-\(d\) is applied and

\[
h_i(x) = \begin{cases} 
\lambda \frac{\sum_{k=0}^d x_{k,2}}{\sum_{k=0}^d x_{k,2}} \mathbb{I}_{\{i \leq d\}}, & \text{if } \sum_{k=0}^d x_{k,2} > 0 \\
(\beta x_{0,1} + x_{d+1,2} \mathbb{I}_{\{i = d\}}) \mathbb{I}_{\{x_{d+1,2} + (d+1) \beta x_{0,1} \leq \lambda\}}, & \text{if } \sum_{k=0}^d x_{k,2} = 0, i \leq d, \\
\frac{x_{i,2}}{y_0} (\lambda - x_{d+1,2} - (d + 1) \beta x_{0,1})^+, & \text{if } \sum_{k=0}^d x_{k,2} = 0, i > d,
\end{cases} \quad (6)
\]

if JBT-\(d\) is applied.

**Theorem 1.** Let \( T < \infty \), \( x^{(0)} \in S_1 \) and assume that \( \|X_N(0) - x^{(0)}\|_w \to 0 \) almost surely. Then, limit points of the stochastic process \((X_N(t))_{t \in [0,T]}\) exist and almost surely satisfy the conditions that define a fluid solution started at \( x^{(0)} \).
Fluid Model and Connection with the Markov Model

**Definition 1.** A continuous function \( x(t) : \mathbb{R}_+ \rightarrow S \) is said to be a fluid model (or fluid solution) if for almost all \( t \in [0, \infty) \)

\[
\begin{align*}
\dot{x}_{0,0} &= \gamma x_{0,2} - \alpha g \mathbb{I}_{\{x_{0,0}>0\}} - \gamma x_{0,2} \mathbb{I}_{\{x_{0,0}=0, \gamma x_{0,2} \leq \alpha g\}} \quad (4a) \\
\dot{x}_{0,1} &= \alpha g \mathbb{I}_{\{x_{0,0}>0\}} - \beta x_{0,1} + \gamma x_{0,2} \mathbb{I}_{\{x_{0,0}=0, \gamma x_{0,2} \leq \alpha g\}} \quad (4b) \\
\dot{x}_{0,2} &= x_{1,2} - h_0(x) + \beta x_{0,1} - \gamma x_{0,2} \quad (4c) \\
\dot{x}_{i,2} &= x_{i+1,2} \mathbb{I}_{\{i \leq B\}} - x_{i,2} + h_{i-1}(x) - h_i(x), \quad (4d)
\end{align*}
\]

where \( g := g(x) : S \rightarrow [0, 1] \), and \( h_i(x) = \min\{\beta x_{0,1}, \lambda\} \) if \( y_0 > 0 \) and otherwise \( (y_0 = 0) \):

\[
h_i(x) = \lambda \frac{y_i^d - y_{i+1}^d}{y_0^d} \quad (5)
\]

if Power-of-d is applied and

\[
h_i(x) = \begin{cases} 
\lambda \frac{x_{i,2}}{\sum_{k=0}^d x_{k,2}} \mathbb{I}_{\{i \leq d\}}, & \text{if } \sum_{k=0}^d x_{k,2} > 0 \\
(\beta x_{0,1} + x_{d+1,2} \mathbb{I}_{\{i = d\}}) \mathbb{I}_{\{x_{d+1,2} + (d+1) \beta x_{0,1} \leq \lambda\}}, & \text{if } \sum_{k=0}^d x_{k,2} = 0, \ i \leq d, \\
\frac{x_{i,2}}{y_0} (\lambda - x_{d+1,2} - (d+1) \beta x_{0,1})^+, & \text{if } \sum_{k=0}^d x_{k,2} = 0, \ i > d,
\end{cases}
\]

if JBT-d is applied.

**Theorem 1.** Let \( T < \infty \), \( x(0) \in S_1 \) and assume that \( \|X^N(0) - x(0)\|_w \rightarrow 0 \) almost surely. Then, limit points of the stochastic process \( (X^N(t))_{t \in [0,T]} \) exist and almost surely satisfy the conditions that define a fluid solution started at \( x(0) \).

---

**Diagram:**

- **Graph:**
  - \( \lambda = 0.7 \)
  - \( \text{Power-of-2} / g(x) = \text{constant} \)
  - 

**Legend:**

- \( \cdot \)
- \( x_{00} \)
- \( x_{01} \)
- \( x_{02} \)
- \( y_1 \)
- \( X_{00} \)
- \( X_{01} \)
- \( X_{02} \)
- \( Y_{01} \)

**Note:**

- **Waste of resources!**

---
Optimal Design

Goal: to design scaling rules ensuring that a global attractor exists and is given by $x^*$ with

$$x^*_{\text{OFF}} = 1 - \lambda, \quad x^*_{\text{ON}} = \lambda$$

(well, $x^*_{0,0} = 1 - \lambda, \quad x^*_{1,2} = \lambda$)

In $x^*$, asymptotic “delay and relative energy optimality” (DREO)
Optimal Design

**Goal**: to design scaling rules ensuring that a global attractor exists and is given by $x^*$ with

$$x^*_{\text{OFF}} = 1 - \lambda, \quad x^*_{1,\text{ON}} = \lambda$$

(well, $x^*_{0,0} = 1 - \lambda$, $x^*_{1,2} = \lambda$)

In $x^*$, asymptotic “delay and relative energy optimality” (DREO)

**Theorem 2**. Let $x(t)$ denote a fluid solution induced by JJQ and any auto-scaling rule $g(x)$ such that

$$g(x) = 0 \text{ if and only if } x_{1,2} + \beta x_{0,1} \geq \lambda.$$

Then, $\lim_{t \to \infty} \|x(t) - x^*\|_w = 0$.

**Theorem 2 (rephrased)**. DREO is obtained *only* by using Join-the-Idle-Queue and a non-zero scale-up rate iff $\lambda > “\text{overall rate at which servers become idle-on”}$.
Empirical Comparison: Synchronous vs Asynchronous

We compare:
- our asynchronous combination of JIQ and Rate-Idle (ALBA), ie, $g(x) = \frac{1}{\lambda}(\lambda - \beta x_{0,1} - x_{1,2})^+$, with
- TABS [Borst et al., 2017], which is synchronous, and achieves DREO.

$a N$ (rate of the auto-scaling clock) set to make both scale-up rates equal
(scale-up rate = number of server initialization signals divided by time horizon)

Our metrics:
- the empirical probability of waiting
- the empirical energy consumption

$$\mathcal{R}_{\text{Wait}} := \frac{p_{\text{Wait}}^{\text{ALBA}}}{p_{\text{Wait}}^{\text{TABS}}}, \quad \mathcal{R}_{\text{Energy}} := \frac{E^{\text{ALBA}}}{E^{\text{TABS}}}$$
Empirical Comparison: Synchronous vs Asynchronous

\[ R_{\text{Wait}} := \frac{p_{\text{ALBA}}^{\text{Wait}}}{p_{\text{TABS}}^{\text{Wait}}}, \quad R_{\text{Energy}} := \frac{E_{\text{ALBA}}}{E_{\text{TABS}}} \]

Possible explanation. Asynchronous is “proactive”: jobs do not necessarily need to wait any time a scale-up decision is taken.