Performance and Stability Analysis of the Task Assignment based on Guessing Size Routing Policy

Josu Doncel

University of the Basque Country, UPV/EHU

joint work with E. Bachmat and H. Sarfati (Ben-Gurion University)

October 22, 2019
Parallel-Server Systems

K homogeneous FIFO queues
Poisson arrivals

Question?
How to balance the load optimally?
Heavy-tailed distribution

A small fraction of jobs make up the half of the load

Example: Bounded Pareto $(1, r, \alpha)$

$$f(s) = \frac{\alpha s^{-\alpha - 1}}{1 - r^{-\alpha}}.$$
Heavy-tailed distribution

A small fraction of jobs make up the half of the load

Example: Bounded Pareto ($\alpha, r, f(s) = \alpha s^{-\alpha - 1} \frac{1}{1 - r^\alpha}$).
A small fraction of jobs make up the half of the load

Example: Bounded Pareto $(1, r, \alpha)$

$$f(s) = \frac{\alpha s^{-\alpha-1}}{1 - r^{-\alpha}}.$$
Known optimality results

JSQ: each incoming job is sent to the server with less number of jobs

Po2: pick $d$ servers at random $\Rightarrow$ JSQ

Disadvantage

Many observations $\Rightarrow$ Not practical
Open-loop Routing Policies

Heavy-tailed: bad performance
- Round-Robin
- Random Splitting
Open-loop Routing Policies

Heavy-tailed: bad performance
- Round-Robin
- Random Splitting

Heavy-tailed: good performance
- SITA: job duration knowledge
Open-loop Routing Policies

Heavy-tailed: bad performance
- Round-Robin
- Random Splitting

Heavy-tailed: good performance
- SITA: job duration knowledge
- Task Assignment with Guessing Size (TAGS)
TAGS Policy

TAGS Policy

Waiting time: $W_1$

---

Waiting time: $W_1$

\[ W_1 = 6 \text{ sec} \]

\[ (0,3] \rightarrow (3,5] \rightarrow (5,8] \rightarrow (8,\infty] \]

---

Waiting time: $W_1$

---

Waiting time: $W_1$

---

Waiting time: $W_1$
TAGS Policy

Waiting time: $W_1$
Waiting time: $W_1 + 3 + W_2 + 5 + W_3$

---

Advantages

- No signaling
- Heavy-tail: good performance
- Durations unknown

---

### Advantages
- No signaling
- Heavy-tail: good performance
- Durations unknown

### Summary of contributions
1. Stability
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis

---

Advantages

- No signaling
- Heavy-tail: good performance
- Durations unknown

Summary of contributions

1. **Stability**
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Proposition

Let \( \rho = \lambda \mathbb{E}[X] \). The TAGS system is stable if and only if

\[
\rho < \frac{\mathbb{E}[X]}{M(X)}
\]

where \( M(X) = \sup_s s(1 - F(s)) \).

\( \Rightarrow \) Critical load \( \rho_{crit}(X) \)
Let $\rho = \lambda \mathbb{E}[X]$. The TAGS system is stable if and only if
\[ \rho < \frac{\mathbb{E}[X]}{M(X)} \]
where $M(X) = \sup_s s(1 - F(s))$.

$\Rightarrow$ Critical load $\rho_{\text{crit}}(X)$

Proposition

Let $X$ be a distribution in $[1, r]$
\[ \rho_{\text{crit}}(X) \leq 1 + \log r. \]
Critical load

**Bounded Pareto \((1, r, \alpha)\)**

If \(\alpha \neq 1\),

\[
\rho_{\text{crit}} = (1 - r^{\alpha-1})(1 - \alpha)^{-1/\alpha}.
\]

If \(\alpha = 1\),

\[
\rho_{\text{crit}} = \frac{r \log r}{r - 1}.
\]
Critical load

Bounded Pareto \((1,r,\alpha)\)

If \(\alpha \neq 1\),

\[
\rho_{\text{crit}} = (1 - r^{\alpha^{-1}})(1 - \alpha)^{-1/\alpha}.
\]

If \(\alpha = 1\)

\[
\rho_{\text{crit}} = \frac{r \log r}{r - 1}.
\]

Tight distribution

\(f(x) = 1/x^2\), for \(x \in [1, r]\)

\(\Rightarrow\) Dirac delta at \(r\) with mass \(r^{-1}\)
Advantages

- No signaling
- Heavy-tail: good performance
- Durations unknown

Summary of contributions

1. Stability
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Proposition

Let $s^{que}$ be the vector of cutoffs that minimizes the maximum mean queue length of the servers. Then, in a system with $h$ hosts,

$$
\mathbb{E}[W(s^{que})] \leq h\mathbb{E}[W^*] + \mathbb{E}[X](h - 1)
$$

$\Rightarrow$ Upper-bound of $\mathbb{E}[W(s^{que})]$. 
Advantages

- No signaling
- Heavy-tail: good performance
- Durations unknown

Summary of contributions

1. Stability
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Bounded Pareto \((1,r,\alpha)\)

**Assumption**

Poisson arrivals to all the servers

\[\Rightarrow \text{Accuracy validation (over-estimation) numerically}\]
Bounded Pareto $(1,r,\alpha)$

**Assumption**

Poisson arrivals to all the servers

$\Rightarrow$ Accuracy validation (over-estimation) numerically

Let $\rho = \lambda \mathbb{E}[X]$. When $r \to \infty$ and $\rho < 1$

**Proposition**

The mean waiting time in a TAGS system with optimal cutoffs is at most two times larger than the mean waiting time of a SITA system with optimal cutoffs.
Bounded Pareto \((1, r, \alpha)\)
Bounded Pareto \((1,r,\alpha)\)

Let \(\rho = \lambda \mathbb{E}[X]\). When \(r \to \infty\) and \(\rho < 1\)

**Proposition**

The mean waiting time in a TAGS system with optimal cutoffs is at most two times larger than the mean waiting time of a SITA system with optimal cutoffs.
Bounded Pareto \((1,r,\alpha)\)

Let \(\rho = \lambda \mathbb{E}[X]\). When \(r \to \infty\) and \(\rho < 1\)

**Proposition**

The mean waiting time in a TAGS system with optimal cutoffs is at most two times larger than the mean waiting time of a SITA system with optimal cutoffs.

Penalty for not knowing the duration of the jobs is at most 2
Bounded Pareto \((1,r,\alpha)\) when \(\rho > 1\)

Big performance difference (stable?)
Bounded Pareto \((1,r,\alpha)\) when \(\rho > 1\)

Big performance difference (stable?)

\[ \tilde{h} = h - i + 1: \text{number of spare servers} \]
\[ \Rightarrow i: \text{minimum number of servers for stability} \]
Bounded Pareto \((1,r,\alpha)\) when \(\rho > 1\)

Big performance difference (stable?)

\[\tilde{h} = h - i + 1: \text{number of spare servers}\]

\[\Rightarrow i: \text{minimum number of servers for stability}\]

Proposition

When \(r \to \infty\) and \(\rho > 1\), the order of magnitude of the optimal mean waiting time of TAGS depends on \(\tilde{h}\) and not on \(h\).
T+W Policy

Josu Doncel (UPV/EHU)

Performance and Stability of TAGS

October 22, 2019 16 / 19
Numerical Experiments with $r < \infty$ and $\rho = 0.5$
Conclusions

Advantages of TAGS

- No signaling
- Heavy-tail: good performance
- Durations knowledge not required

Our contributions

1. Stability
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Conclusions

Advantages of TAGS

- No signaling
- Heavy-tail: good performance
- Durations knowledge not required

Summary of contributions

1. Stability $\Rightarrow$ Critical load
2. Optimal performance
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Conclusions

Advantages of TAGS

- No signaling
- Heavy-tail: good performance
- Durations knowledge not required

Our contributions

1. Stability $\Rightarrow$ Critical load
2. Optimal performance $\Rightarrow$ Bounds
3. (Bounded) Pareto distribution
   - Comparison with SITA
   - Optimal performance analysis
Conclusions

Advantages of TAGS
- No signaling
- Heavy-tail: good performance
- Durations knowledge not required

Our contributions
1. Stability $\Rightarrow$ Critical load
2. Optimal performance $\Rightarrow$ Bounds
3. (Bounded) Pareto distribution
   - Comparison with SITA $\Rightarrow$ when $\rho < 1$
   - Optimal performance analysis $\Rightarrow$ when $\rho > 1$
Future Research

TODOs in our model

Non-asymptotic analysis and comparison with other policies
Future Research

TODOs in our model
Non-asymptotic analysis and comparison with other policies

Extensions to energy networks
EPN
On-off servers
Future Research

**TODOs in our model**
- Non-asymptotic analysis and comparison with other policies

**Extensions to energy networks**
- EPN
- On-off servers

THANKS