SAPIENZA
UNIVERSITÀ DI ROMA

NBD Homework 2

Group: Erlang

Dario Baraghini
Michelangelo Saveriano
Paolo Falcone
Nicola Calabrese

May 2021

SITA Policy

Assuming that the jobs arrive at the dispatcher according to a Poisson process with mean rate Λ and for the *reproducibility property* of the Poisson random variables:

$$\lambda_1 = P(L \leq \theta)\Lambda = \left(1 - \left(\frac{b}{\theta}\right)^{\alpha}\right)\Lambda, \qquad \lambda_2 = P(L > \theta)\Lambda = \left(\frac{b}{\theta}\right)^{\alpha}\Lambda$$

Then we know that the workload L is distributed following a Pareto random variable, conditioning on θ it turns out that the two server service times are still Pareto random variables:

$$P(X_1 \leq x) = \frac{1 - b^{\alpha}(x\mu_1)^{-\alpha}}{1 - \left(\frac{b}{\theta}\right)^{\alpha}} 1_{[b,\theta]}(x\mu_1), \qquad P(X_2 \leq x) = 1 - \left(\frac{\theta}{x\mu_2}\right)^{\alpha} 1_{[\theta,\infty]}(x\mu_2)$$

For the SITA algorithm the dispatching delay is 0 because it doesn't require any message from the servers (just checks workload dimension), $\mathbb{E}[D] = \mathbb{E}[S]$ where $\mathbb{E}[S]$ for this situation is:

$$\mathbb{E}[S] = \mathbb{E}[X] + \frac{\lambda \mathbb{E}[X^2]}{2(1 - \lambda \mathbb{E}[X])}$$

Imposing the constraints on stability conditions, the optimization problem becomes:

$$\begin{aligned} \min_{\theta} \quad & \left(1 - \left(\frac{b}{\theta}\right)^{\alpha}\right)\mathbb{E}[S_1] + \left(\frac{b}{\theta}\right)^{\alpha}\mathbb{E}[S_2] \\ s.t. \quad & \lambda_1 \mathbb{E}[L_1] < \mu_1 \qquad \text{where } L_1 \sim BoundedPareto(b, \theta, \alpha) \\ & \lambda_2 \mathbb{E}[L_2] < \mu_2 \qquad \text{where } L_2 \sim Pareto(\theta, \alpha) \end{aligned}$$

Solving the two constraints above we obtain the interval where the parameter θ can vary:

$$max\left\{b, \left(\frac{\alpha\Lambda b^{\alpha}}{\mu_2(\alpha - 1)}\right)^{\frac{1}{\alpha - 1}}\right\} < \theta < \left(\frac{\alpha\Lambda b^{\alpha}}{\alpha\Lambda b - \mu_1(\alpha - 1)}\right)^{\frac{1}{\alpha - 1}}$$

Random Policy

In the random dispatching policy the job dispatcher decides to send a job to server 1 with probability p and to server 2 with probability $1 - p$. From this derives that:

$$\lambda_1 = p\Lambda, \qquad \lambda_2 = (1 - p)\Lambda$$

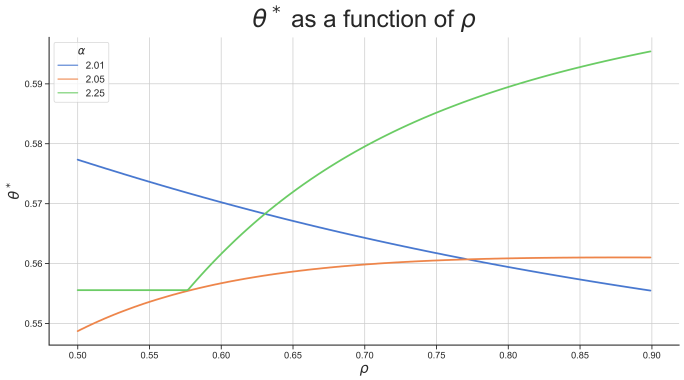
Since the dispatching process does not imply any filtering based on the workload the 2 service time distributions are also *Pareto distributions*. The expected values for the service times of the servers are:

$$\mathbb{E}[X_1] = \frac{1}{\mu_1}\frac{\alpha b}{\alpha - 1}, \qquad \mathbb{E}[X_2] = \frac{1}{\mu_2}\frac{\alpha b}{\alpha - 1}$$

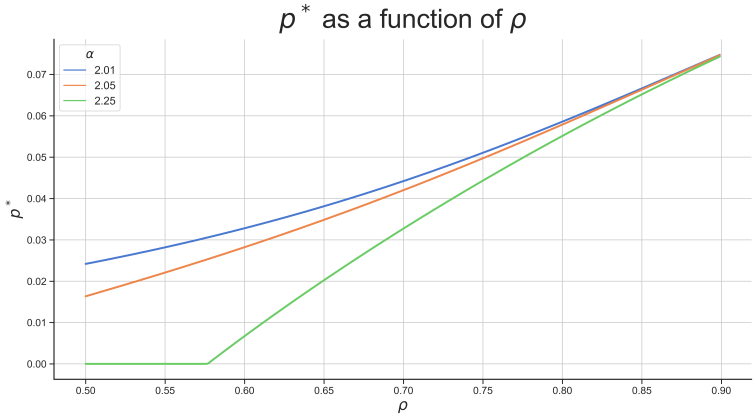
Also for the Random policy the dispatching delay is 0, thus the mean delay is reduced to the service time S.The optimization problem is:

$$\begin{aligned} \min_{p} \quad & p\mathbb{E}[S_1] + (1 - p)\mathbb{E}[S_2] \\ s.t. \quad & 1 - \frac{1}{\Lambda \mathbb{E}[X_2]} < p < \frac{1}{\Lambda \mathbb{E}[X_1]}, \qquad 0 \leq p \leq 1 \end{aligned}$$

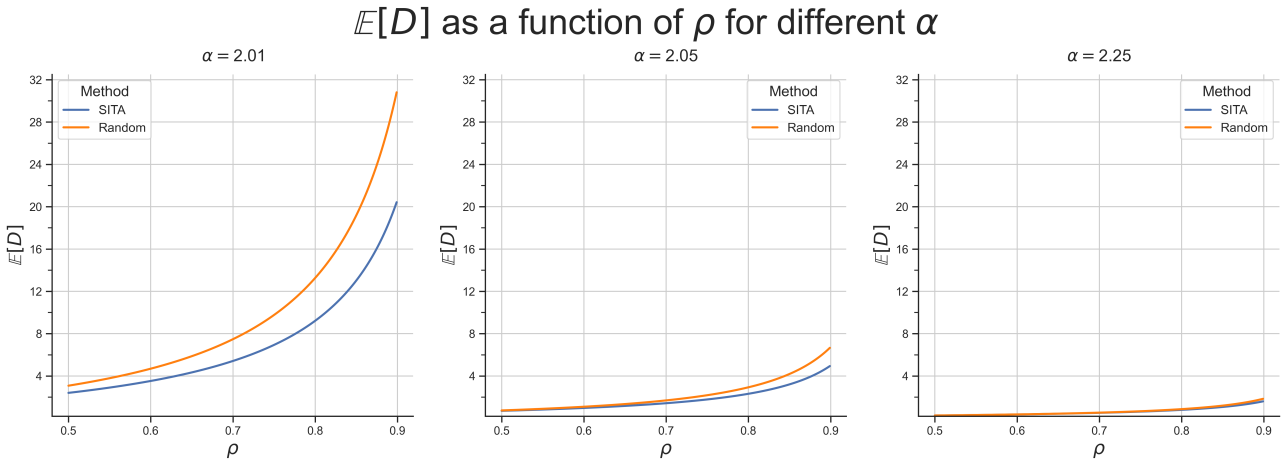
where the constraints derive from the usual stability conditions.



Since the workload L is distributed like a Pareto r.v., with expectation $E[L] = \frac{\alpha b}{\alpha - 1}$, increasing α means that the mean workload decreases and more work can be dispatched to server 1. In fact for $\alpha = 2.01$ (blue line) the mean workload is very high and it's convenient to use mainly server 2, which has more processing capability. For the green line ,instead, we can see an interesting behaviour: when ρ is small enough θ^* coincides with the lower bound b, this means that no job goes towards server 1.



In this figure we can see a peculiar graph tendency: it seems that, when $\rho \to 1$, the optimal probability of dispatching workloads between servers is equal to the ratio between the servers' processing capacity and the system capacity, $p = \frac{\mu_1}{\mu_1 + \mu_2}$, $1 - p = \frac{\mu_2}{\mu_1 + \mu_2}$.



In the FCFS scheduling policy the expected system time is : $\mathbb{E}[S] = \mathbb{E}[X] + \frac{\rho \mathbb{E}[X]}{1-\rho} \frac{1+C_{X^2}}{2}$, where C_{X^2} is the squared coefficient of variation and it measures the variability of the service time. Thus, for the expression of Pareto's variance, if α increases the variance of the workload L decreases, and consequently also the total mean delay $\mathbb{E}[D]$. Moreover as we can see the SITA policy is better than Random for smaller α but when α grows we can't find an overall best policy.