Distributed Simulation of Supercomputer Model with Heavy Tails

Alexander Golovin, Alexander Rumyantsev, Sergey Astafiev

Institute of Applied Mathematical Research Karelian Research Centre, RAS,
Petrozavodsk State University, Petrozavodsk, Russia.

The publication has been prepared with the support of Russian Science Foundation according to the research project No.21-71-10135 https://rscf.ru/en/project/21-71-10135/

Russian Supercomputing Days, 2022

Introduction

The simultaneous service multiserver queue had recently experienced a spike of interest due to a new application in the context of supercomputers (supercomputer model).
Due to a rather sophisticated machinery, the model permits analysis only in rather restrictive cases such as exponential interarrival/service times, or a small number of servers.
Thus, in realistic environment, the study needs to be performed numerically e.g. by simulation, whereas the known analytical solutions can be used for simulation model validation.
One of the specific observations based on the analysis of the supercomputer workload logs are heavy-tailed distributions of the key random variables such as interarrival/service times. It is well known that the presence of heavy tails may dramatically affect the system performance.
Using regenerative modeling in a distributed computing system, we considered the impact of heavy-tailed distribution on the performance of supercomputer model.

Discrete-Event Simulation (DES)

Generalized Semi-Markov Processes: $\Theta = \{\boldsymbol{X}(t), \boldsymbol{T}(t)\}_{t \geq 0}$
If the process Θ is regenerative, i.e. there exists some sequence of the time epochs $\{\tau_k\}_{k \geq 1}$ with iid time intervals, such that the random elements $\{\boldsymbol{X}(t), \boldsymbol{T}(t)\}_{\tau_i \leq t < \tau_{i+1}}$ are iid, then the steady-state (time average) performance estimate is obtained by using the integral estimate on the corresponding regeneration cycles,

$$Y_j = \sum_{k=\beta_{j-1}}^{\beta_j - 1} \chi(\boldsymbol{X}^{(k)})(t_{k+1} - t_k), \quad j \geq 1$$

where $\{t_k\}_{k \geq 0}$ is the sequence of event epochs, $\boldsymbol{X}^{(k)} = \boldsymbol{X}(t_k)$, while the regeneration epochs in discrete time β_j (starting from $\beta_0 = 0$) are obtained from

$$t_{\beta_j} = \tau_j, \quad j \geq 1$$

Performance estimation of the model in steady state

The pointwise estimate is then obtained as

$$\overline{r}_n = \frac{Y_1 + \cdots + Y_n}{\tau_n} = \frac{S_{Y_n}}{S_{R_n}} \tag{1}$$

Assuming $\mathrm{E}(Y_1 + \tau_1)^2 < \infty$, regenerative central limit theorem allows one to obtain $(1 - 2\gamma)\%$ confidence estimate for the true mean $r = \mathrm{E}Y_1/\mathrm{E}\tau_1$ as follows:

$$\overline{r}_n \pm \frac{h_\gamma \sqrt{\overline{\mathrm{Var}}(n)}}{\sqrt{n} S_{R_n}} \tag{2}$$

unbiased estimator $\overline{\mathrm{Var}}(n)$ is obtained as

$$\overline{\mathrm{Var}}(n) = \frac{nS_{Y_n^2} - (S_{Y_n})^2 - 2\overline{r}_n(nS_{YR_n} - S_{Y_n}S_{R_n}) + \overline{r_n}^2(nS_{R_n^2} - (S_{R_n})^2)}{n(n-1)}$$

Speedup computing

Due to additive way of computations in (1) and (2) which are based on partial sums (3) and (4) of iid elements, it is natural to use the time-parallel computations.

$$S_{Y_n} = Y_1 + \cdots + Y_n, \quad S_{R_n} = R_1 + \cdots + R_n = \tau_n, \quad n \geq 1 \tag{3}$$

$$S_{Y_n^2} = Y_1^2 + \cdots + Y_n^2, \quad S_{YR_n} = Y_1 R_1 + \cdots + Y_n R_n, \quad S_{R_n^2} = R_1^2 + \cdots + R_n^2. \tag{4}$$

We use the `simulato`[1] package for R language to implement the model and the `RBOINC`[2] R package to conduct the experiments on a self-hosted BOINC desktop grid facility. Both packages are available at `R-Forge` package distribution system.

[1] `https://r-forge.r-project.org/projects/simulato`

[2] `https://r-forge.r-project.org/projects/rboinc`

Heavy-tailed Distribution Sampling

An important special case of a heavy-tailed distribution is Pareto distribution.

$$F(x) = P(X \leq x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha}, \quad x > x_0.$$

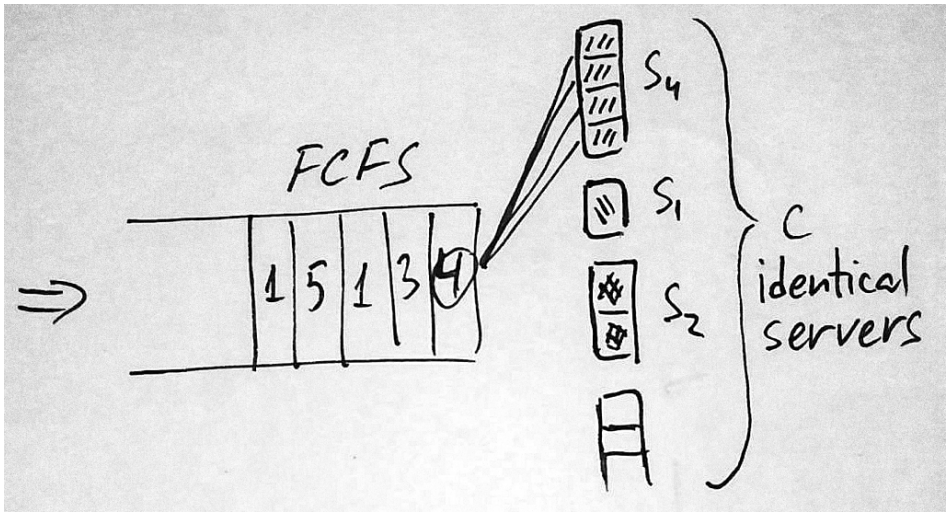
Using inverse function method for generation random value we obtain:

$$x = \frac{x_0}{\sqrt[\alpha]{u}},$$

where u is a random variable uniformly distributed on $(0, 1]$.
The problems with using such distributions in simulations:

- The system is in transient mode for the entire duration of the simulation, no matter how long it lasts.
- A very large number can be generated, which is unattainable in practice. For example, in real data networks, the maximum packet size is always upper bounded by some value. Thus, when generating, we may need to limit the maximum possible value.

GSMP model of a supercomputer



- c identical servers receiving input of customers of c classes
- (random) class N_i of a customer i = number of servers required to serve the customer
- all servers dispatched to the customer are seized and released simultaneously for the same random amount of time S_i

The distinctive feature of the supercomputer model compared to the classical multiserver queue is the simultaneous service of a customer by a random number of servers, which makes the discipline non-work-conserving and complicates the analysis.

Architecture

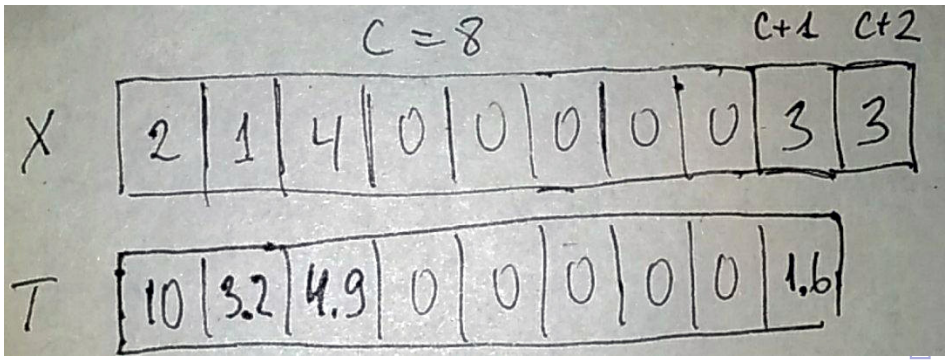
The state vector with $c+2$ components $X_1, \ldots, X_{c+2}$ where

- $X_1, \ldots, X_c$ are the classes of not more than c customers being served (some elements may be zero),
- X_{c+1} is the class of the customer at the head of the queue,
- X_{c+2} number of customers in the queue (after the first one, if any).

The timers:

- $T_1, \ldots, T_c$ are the remaining service times of the customers being served (corresponding to the nonzero state components),
- T_{c+1} is the remaining interarrival time.

It remains to note that the timer speed is unit for the nonzero timers.



PASTA property

$\boldsymbol{W}_\infty$ is the discrete-time limit

$$\boldsymbol{W}_{i+1} = \mathrm{R}\left(\boldsymbol{e}_{1:N_i}(W_{i,N_i} + S_i) + (\mathbf{1} - \boldsymbol{e}_{1:N_i}) \circ \boldsymbol{W}_i - \mathbf{1}\tau_i\right)^+$$

$\boldsymbol{W}(\infty)$ is the continuous-time

$$\boldsymbol{W}(t) = \mathrm{R}\left(\boldsymbol{e}_{1:N_i}(W_{A(t),N_{A(t)}} + S_{A(t)}) + (\mathbf{1} - \boldsymbol{e}_{1:N_{A(t)}}) \circ \boldsymbol{W}_{A(t)} - \mathbf{1}\bar{\tau}(t)\right)^+$$

$$\boldsymbol{W}(\infty) =_d \mathrm{R}\left(\boldsymbol{e}_{1:N}(W_{\infty,N} + S + (\mathbf{1} - \boldsymbol{e}_{1:N}) \circ \boldsymbol{W}_\infty - \mathbf{1}\tau\right)^+ =_d \boldsymbol{W}_\infty$$

where $(\cdot)^+ = \max(0, \cdot)$; R orders the components ascendingly; $\circ$ is Hadamard (componentwise) product; $\boldsymbol{e}_{1:N_i}$ is the vector with N_i (first) unit components; $\bar{\tau}(t) = t - \tau_{A(t)}$ is the time since most recent arrival; $A(t)$ is the counting process giving the number of most recent arrival before time $t \geq 0$

PASTA Inequality

A random variable τ with distribution function F_τ is new-better-than-used (NBU) if for any $x, y \geq 0$ the following inequality holds good,

$$\overline{F}_T(x+y) \leq \overline{F}_T(y)\overline{F}_T(x)$$

Define the remaining interarrival time

$$\hat{\tau}(t) = t_{A(t)+1} - t$$

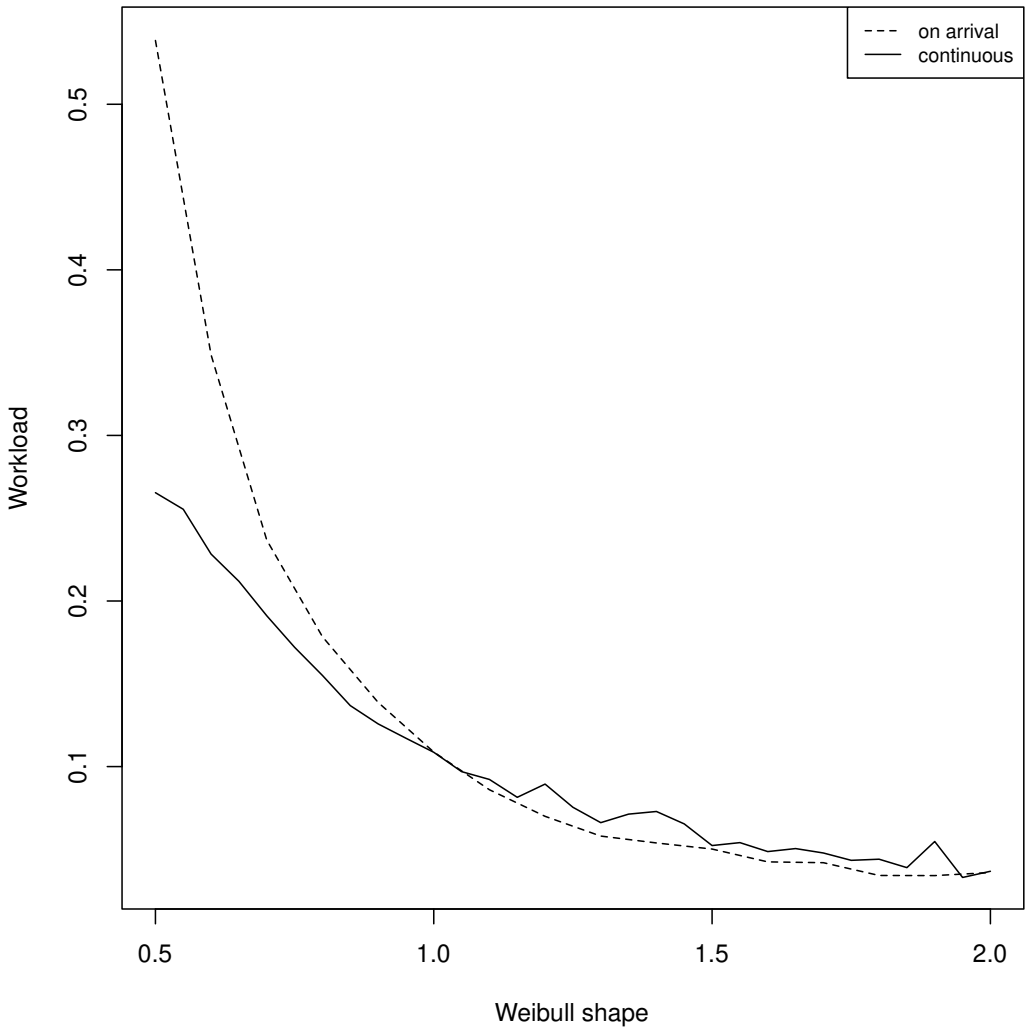
PASTA Inequality:

$$\begin{aligned} \boldsymbol{W}(\infty) &=_d \mathrm{R}\left(\boldsymbol{e}_{1:N}(W_{\infty,N} + S + (\mathbf{1} - \boldsymbol{e}_{1:N}) \circ \boldsymbol{W}_\infty - \mathbf{1}\tilde{\tau}\right)^+ \\ &\geq_d \mathrm{R}\left(\boldsymbol{e}_{1:N}(W_{\infty,N} + S + (\mathbf{1} - \boldsymbol{e}_{1:N}) \circ \boldsymbol{W}_\infty - \mathbf{1}\tau\right)^+ =_d \boldsymbol{W}_\infty \end{aligned} \quad (5)$$

A specific example when the NBU/NWU property holds in parametric way is the Weibull distribution

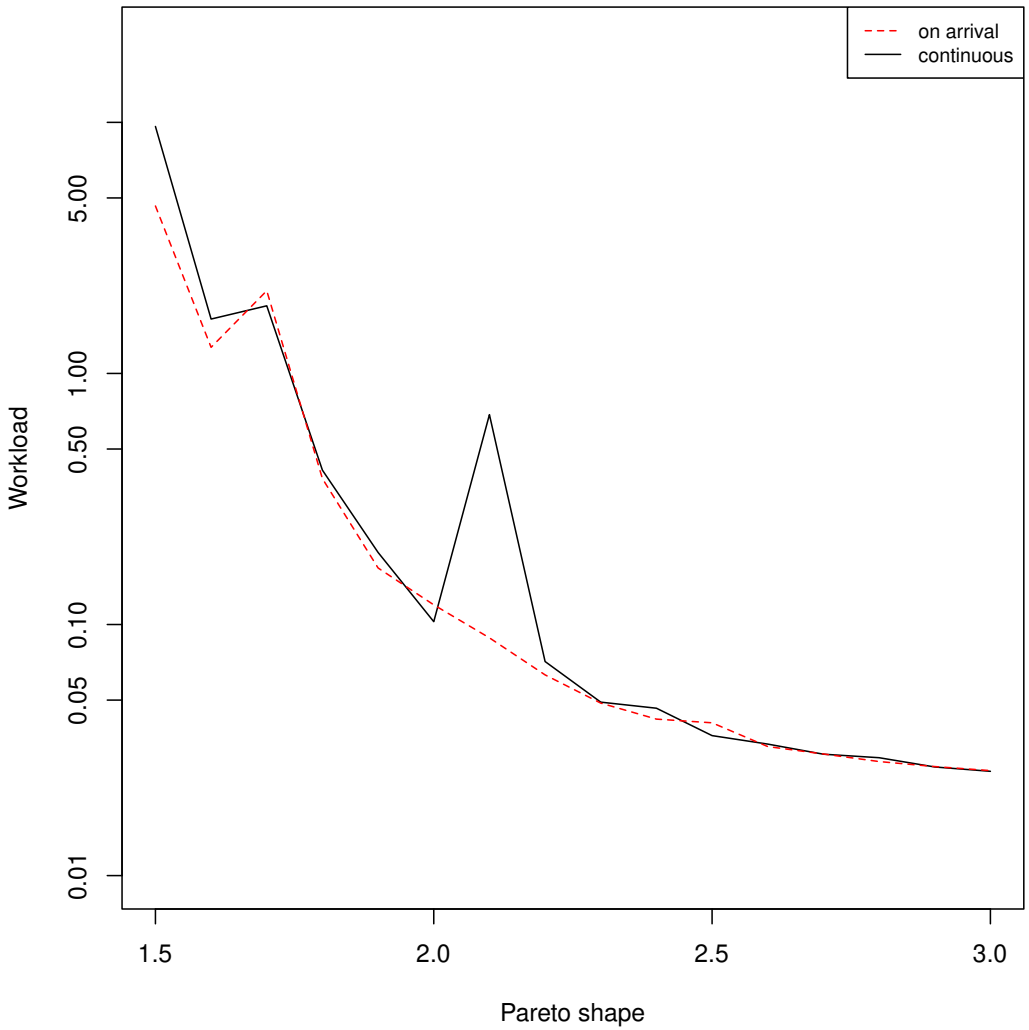
$$\mathrm{P}\{\tau \leq x\} = 1 - e^{-(x/b)^a}, \quad a, b > 0, x \geq 0$$

Numerical Experiments. PASTA Inequality



The estimates
of the mean stationary summary workload in continuous time, $W(\infty)$, and at arrival epochs, W_∞, in Weibull/Pareto/100 cluster model.

Numerical Experiments. Estimator Efficiency



The estimates of the mean stationary waiting time in continuous time, $W(\infty)$, and at arrival epochs, W_∞, vs. Pareto shape α in M/G/100-type supercomputer model.

Conclusions and Future Work

- Some important aspects of a novel supercomputer model were demonstrated by using simulation and heavy-tailed distributions:
  - the PASTA inequality was illustrated
  - direct/indirect waiting time estimator efficiency was compared
- Numerical experiments were performed

Future plans:

- Deeply investigate the effects of heavy tails on system performance for a system of increasing size and/or under increasing load

Thank you for attention!

Golovin Alexander
Institute of Applied Mathematical Research
Karelian Research Centre RAS
golovin@krc.karelia.ru
ResearcherID: AAG-2341-2021
ORCID: orcid.org/0000-0003-1325-3739
ScopusID: 56902328200

ResearchGate: https://www.researchgate.net/profile/Alexander-Golovin