Heavy-traffic Universality of Redundancy Systems with Assignment Constraints

Ellen Cardinaels¹, Sem Borst¹ and Johan van Leeuwaarden²

¹TU/e (NL), ²Tilburg University (NL), @cardinaels@tue.nl

One sentence summary: The parallel-server system with redundancy scheduling achieves full resource pooling in the heavy-traffic regime and exhibits strong insensitivity to the underlying assignment constraints.

Keywords: heterogeneity, parallel-server systems, redundancy scheduling, heavy-traffic limit, state space collapse, resource pooling.

1 Model description

Parallel-server system

- $N$ servers with speeds $\mu_1, \ldots, \mu_N$.
- Assignment constraints: Type-$s$ jobs can only be served by servers $S \subseteq \{1, \ldots, N\}$.
- Type-$s$ jobs arrive according to Poisson process with rate $\lambda_p$.
- Service requirements are exponentially distributed with mean $1/\mu$.

Redundancy scheduling

- Cancel-on-completion (c.o.c.)
- i.i.d. replicas
- FCFS service discipline

2 Intermediate result

Proposition 1

If the stability conditions [1] hold, the joint PGF of the number of jobs of each type for the redundancy c.o.c. policy is given by

$$E \left[ \prod_{j=1}^{N} Q_{j}^{q_{j}} \right] = f(q) / f(1)$$

where $q$ and $1$ are $|\mathcal{S}|$-dimensional vectors with entries $q_j \leq 1$ and

$$f(q) = \prod_{j=1}^{N} \left( 1 + \sum_{i=1}^{m_{ij}} \sum_{j=1}^{m_{ij}} \lambda p_{ij} x_i \right) \times \ldots \times \prod_{j=1}^{N} \left( 1 - \frac{\lambda}{\mu(S_{1,\ldots,j})} \sum_{i=1}^{m_{ij}} p_{ij} x_i \right)^{-1}.$$  

The $m$-dimensional vector $S$ consists of $m$ different job types, and the set consisting of all these vectors is denoted by $\mathcal{S}^{m}$.

3 Heavy-traffic result

Theorem 1 (State space collapse and complete resource pooling)

If the CRP conditions [2] hold, then

$$\left( 1 - \frac{\lambda}{\mu} \right) Q_{S} \overset{d}{\rightarrow} \text{Exp}(1) \text{ (response time)}$$

as $\lambda \uparrow \mu$.

Notation

- $\mathcal{S}$, set of all job types;
- $\lambda$, total arrival rate;
- $\mu$, total service rate of the system;
- $p_{ij}$, fraction of jobs with type $i$ in state $j$;
- $\mu_\mathcal{S}$, aggregate service rate of servers compatible with the job types in $\mathcal{S}$.

Proof sketch - Proposition 1

Substitute $x_i = \exp \left( - \left( 1 - \frac{\lambda}{\mu} \right) t Q_{ij} \right)$. Convergence of the MGFs:

$$E \left[ \exp \left( - \left( 1 - \frac{\lambda}{\mu} \right) \sum_{S \in \mathcal{S}} \lambda_{S} t Q_{S} \right) \right] \rightarrow (1 + \sum_{S \in \mathcal{S}} p_{S} t Q_{S})^{-1}.$$  

Corollaries

- $\left( 1 - \frac{\lambda}{\mu} \right) W_j \overset{d}{\rightarrow} \text{Exp}(1)$ (waiting time)
- $\left( 1 - \frac{\lambda}{\mu} \right) V_j \overset{d}{\rightarrow} \text{Exp}(1)$ (response time)

Supplementary results

- Redundancy cancel-on-start (c.o.s.)
- Convergence of the moments

$$\lim_{\lambda \uparrow \mu} E \left[ \left( 1 - \frac{\lambda}{\mu} \right) Q_{S} \right] = n p_{S}$$

• Relaxed CRP condition

Department of mathematics and computer science