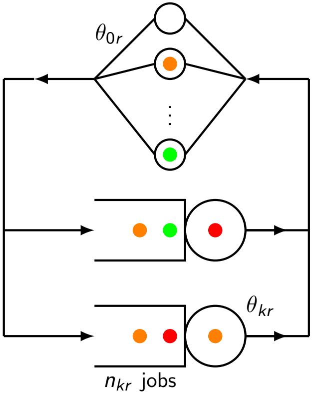
Imperial College London

Facilitating Load-Dependent Queueing Analysis Through Factorization

Giuliano Casale, Peter G. Harrison, Hong Wai Ong
Imperial College London

IFIP PERFORMANCE 2021

Closed load-independent (LI) QNs

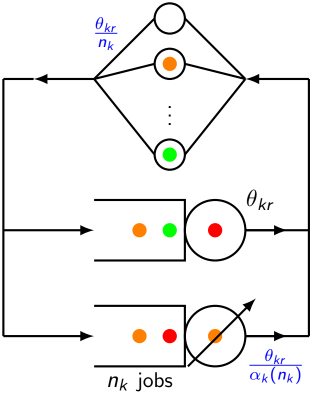


M stations
R classes
N jobs, N_r in class r, $\boldsymbol{N} = (N_r)$

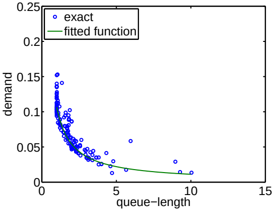
Product-form solution:

$$\pi(\boldsymbol{n}) = \frac{1}{G_{\boldsymbol{\theta}}(\boldsymbol{N})} \prod_{r=1}^{R} \frac{\theta_{0r}^{n_{0r}}}{n_{0r}!} \prod_{k=1}^{M} n_k! \prod_{r=1}^{R} \frac{\theta_{kr}^{n_{kr}}}{n_{kr}!}$$

Closed load-dependent (LD) QNs



Load-dependent demand scalings $\alpha_k(n_k)$



Product-form solution:

$$\pi(\boldsymbol{n}) = \frac{1}{H_{\boldsymbol{\theta}}(\boldsymbol{N})} \prod_{k=1}^{M} \frac{n_k!}{\alpha_k(n_k)} \prod_{r=1}^{R} \frac{\theta_{k,r}^{n_{k,r}}}{n_{k,r}!}$$

$H_{\boldsymbol{\theta}}(\boldsymbol{N})$ also enables performance metric computation.

Related work

Solving a load dependent (LD) QN model:

- MVA-LD: load-dependent mean-value analysis
- CA: Load-dependent convolution algorithm
- RECAL: Load-dependent RECAL method
- AMVA: Queue-dependent approximate MVA
- ODEs: mean-field approximation for multi-server stations
- ...

Exact methods $O(N^{1+\min(M,R)})$ in time and space

Approximations can be unstable, feature low accuracy, or work in special cases only (e.g., multi-server stations).

Key contributions

We show that if multiclass service demands are load-dependent up to a finite population limit (limited load-dependence), then:

- Exact solutions factorize into the products of two terms:
  1. a factor obtained by solving a model without load-dependence
  2. a factor obtained by solving a load-dependent model on a reduced state space
- The second factor may be effectively approximated using simpler single-class LD models.

We then develop novel exact and approximate algorithms that leverage these properties.

Explicit form for load-independent models

Single-class load-independent (LI) models can be solved explicitly in $O(1)$ with respect to the number of jobs N.

Let $g_{\boldsymbol{\theta}}(N)$ be the single class normalizing constant. If demands are non-identical then

$$g_{\boldsymbol{\theta}}(N) = \sum_{i=1}^{M} \frac{\theta_i^{N+M-1}}{\prod_{k\neq i}(\theta_i - \theta_k)}$$

$\rightarrow$ How about load-dependent (LD) models?

Explicit form for multi-server models

Gordon (OPRE'90) obtains for multi-server models:

$$h_{\boldsymbol{\theta}}(N) = \sum_{0 \leq \boldsymbol{v} < \boldsymbol{s}} \sum_{i=1}^{M} \frac{\sigma_i^{N+M-v-1}}{\prod_{j \neq i}(\sigma_i - \sigma_j)} \left(\prod_{k=1}^{M} \frac{\theta_k^{v_k}}{v_k!} \left(1 - \frac{v_k}{s_k}\right)\right)$$

where we define the scaled demands $\sigma_i = \theta_i / s_i$ and $\boldsymbol{\sigma} = (\sigma_i)$, s_i being the number of servers in node i.

Multi-server models are a special case of limited load-dependent (LLD) models:

$$\exists s_k \text{ s.t. } \alpha_k(n_k) = const, \qquad \forall n_k \geq s_k$$

The results generalizes to LLD models if we set $\sigma_i = \theta_i / \alpha_i(s_i)$.

Single-class LLD: our solution

Let $h_{\boldsymbol{\theta}}(N)$ be the single class LLD normalizing constant. We find:

$$h_{\boldsymbol{\theta}}(N) = \sum_{0 \le \boldsymbol{v} < \boldsymbol{s}} g_{\boldsymbol{\sigma}}(N - v)\Phi_{\boldsymbol{\theta}}(\boldsymbol{v})$$

where $\Phi_{\boldsymbol{\theta}}(\boldsymbol{v}) = \prod_{k=1}^{M} \phi_k(v_k)$, in which

$$\phi_k(v_k) = \begin{cases} \dfrac{\theta_k^{v_k}}{\prod_{j=1}^{v_k} \alpha_k(j)} \left(1 - \dfrac{\alpha_k(v_k)}{\alpha_k(s_k)}\right) & \text{if } v_k > 0 \\ 1 & \text{otherwise} \end{cases}$$

We also find asymptotic expressions as $N \to \infty$ (cf. paper).

Multiclass LLD models

We show that the multiclass normalizing constant is obtained from the single-class one by finite differences, i.e.,

$$H_{\boldsymbol{\theta}}(\boldsymbol{N}) = \sum_{0 \le \boldsymbol{n} \le \boldsymbol{N}} \frac{(-1)^{N-n}}{N_1! \cdots N_R!} \prod_{r=1}^{R} \binom{N_r}{n_r} h_{\boldsymbol{\theta}\boldsymbol{n}}(N)$$

where $\boldsymbol{n} = (n_1, \ldots, n_R)^T$. Plugging the explicit form of $h_{\boldsymbol{\theta}\boldsymbol{n}}(N)$, we find the following factorization:

$$\underbrace{H_{\boldsymbol{\theta}}(\boldsymbol{N})}_{\text{LD norm. const.}} = \underbrace{\Gamma(\boldsymbol{N})}_{\text{correction factor}} \cdot \underbrace{G_{\boldsymbol{\sigma}}(\boldsymbol{N})}_{\text{LI norm. const.}}$$

Imperial College London

LLD correction factor

The *LLD correction factor* $\Gamma(\boldsymbol{N})$ is the quantity

$$\Gamma(\boldsymbol{N}) = \sum_{v=0}^{V} \sum_{\substack{\boldsymbol{d}\geq 0:\\ |\boldsymbol{d}|=v}} \prod_{(\boldsymbol{s},r)\in P(\boldsymbol{d},\boldsymbol{N})} X_r^{\boldsymbol{\sigma}}(\boldsymbol{s}) E_{\boldsymbol{\theta}}(\boldsymbol{d})$$

Here, $X_r^{\boldsymbol{\sigma}}(\boldsymbol{N})$ is the class-r throughput in a LI model with demands $\boldsymbol{\sigma}$ and P is a sequence of population vectors.

$E_{\boldsymbol{\theta}}(\boldsymbol{d})$ is a LD normalizing constant for a model with at most $V = \min(N, \sum_{k=1}^{M}(s_k - 1))$ jobs.

Integral forms

We obtain general formulas for LD normalizing constants also applicable to computing $E_{\boldsymbol{\theta}}(\boldsymbol{d})$.

Since the normalizing constant $H(\boldsymbol{N})$ is a finite difference, the Norlund-Rice theorem gives after manipulations

$$H_{\boldsymbol{\theta}}(\boldsymbol{N}) = \frac{1}{(2\pi)^R} \int_0^{2\pi} \cdots \int_0^{2\pi} \Re h_{\Theta(\boldsymbol{t}-\boldsymbol{\beta}^T\boldsymbol{t})}(N) d\boldsymbol{t}$$

where $\beta = \boldsymbol{N}/N$, $\Theta(\boldsymbol{t}) = \boldsymbol{\theta} \cdot (e^{it_1}, \ldots, e^{it_R})^T$, and the integrand is thus a normalizing constant with complex demands.

Formulas for the derivatives of $\Re h$ and $\Im h$ are found to compute Laplace-type approximations of the above integral.

Reduction heuristic (RD)

Alternatively, the normalizing constant may be approximated as

$$H_{\boldsymbol{\theta}}(\boldsymbol{N}) \approx \gamma(\boldsymbol{N}) G_{\boldsymbol{\sigma}}(\boldsymbol{N})$$

where

$$\gamma(\boldsymbol{N}) = \sum_{v=0}^{V} \left(\frac{N - (v-1)^{+}}{N} \right) e_{\boldsymbol{\rho}}(v)$$

where $(v-1)^{+} = \max(0, v-1)$, $\boldsymbol{\rho} = \boldsymbol{\theta} \boldsymbol{X}^{\boldsymbol{\sigma}}(\boldsymbol{N})$, and $e_{\boldsymbol{\rho}}(v)$ is a single class LD normalizing constant.

Reduction heuristic (RD)

Reduction heuristic (RD) translates this result to mean-values:

$$X_r(\boldsymbol{N}) \approx \frac{\gamma(\boldsymbol{N} - 1_r)}{\gamma(\boldsymbol{N})} X_r^{\boldsymbol{\sigma}}(\boldsymbol{N})$$

The $\gamma(N)$ scaling factor can be computed with our explicit formulas or with asymptotic expansions.

RD heuristic validation:
$\rightarrow$ 1%-6% mean absolute relative error on thousands of models
$\rightarrow$ Shown typically more accurate than AMVA and fluid ODEs.

Imperial College London

Conclusion

Main achievements:

- Exact explicit solution for single-class LLD models
- Factorized solution of multi-class LLD models
- Integral forms for multi-class LLD models (more in the paper)
- Mean-value analysis approximation (RD heuristic)

Further results in the paper:

- Detailed numerical results
- Applications to response time distribution analysis
- Applications to non-product form model approximation

Possible lines for future work:

- Class-dependent scalings
- Whittle networks with closed populations