Machine Learning Based Analysis of Queueing Systems

Siamak Khayyati
Assistant Professor, HEC Liege, The Management School of the University of Liege

May 20, 2025

Motivation

- Complex manufacturing systems suffer from long cycle times
  - In automative manufacturing lead times can vary from few weeks to months
    - To improve the service level, cycle times need to be minimized and the lead times need to be picked properly

The structure of the paint shop in the Ford Otosan automotive production plant

Motivation

- To solve this optimization problem, the **distribution** of $CT(x)$ needs to be determined/approximated reliably
- This can be achieved by modeling the system as a **queueing network**

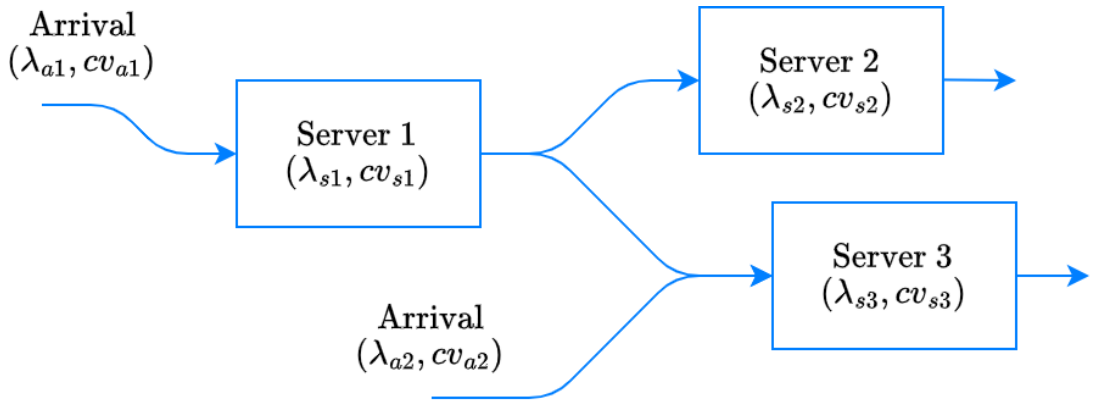
$$\min_{x \in X} \quad C(x)$$

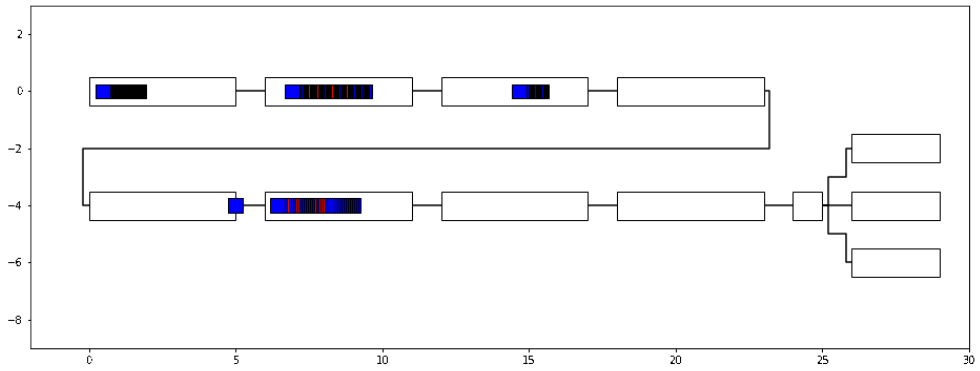
$$\text{subject to} \quad \mathbb{P}\left[CT(x) \leq L_d\right] \geq \alpha$$

Notation	Description
x	The design variable of interest for the production system including the number of servers in each station, the service discipline, …
$C(x)$	The total cost of design x
$CT(x)$	The total cycle time for design x
L_d	The lead time
α	The service level.

Queueing Network Analysis

- How complex queueing systems can be analyzed?
  - Exact solution
    - State space explosion
      - 1000 steps → size of state space: N^{1000}
  - Simulation
    - Too slow for optimization
    - Could be as slow as the real system!
  - Decomposition





Queueing Network Analysis

- Decomposition (Kuehn 1979)
  - Delay ($u = \frac{\lambda_a}{\lambda_s}$):

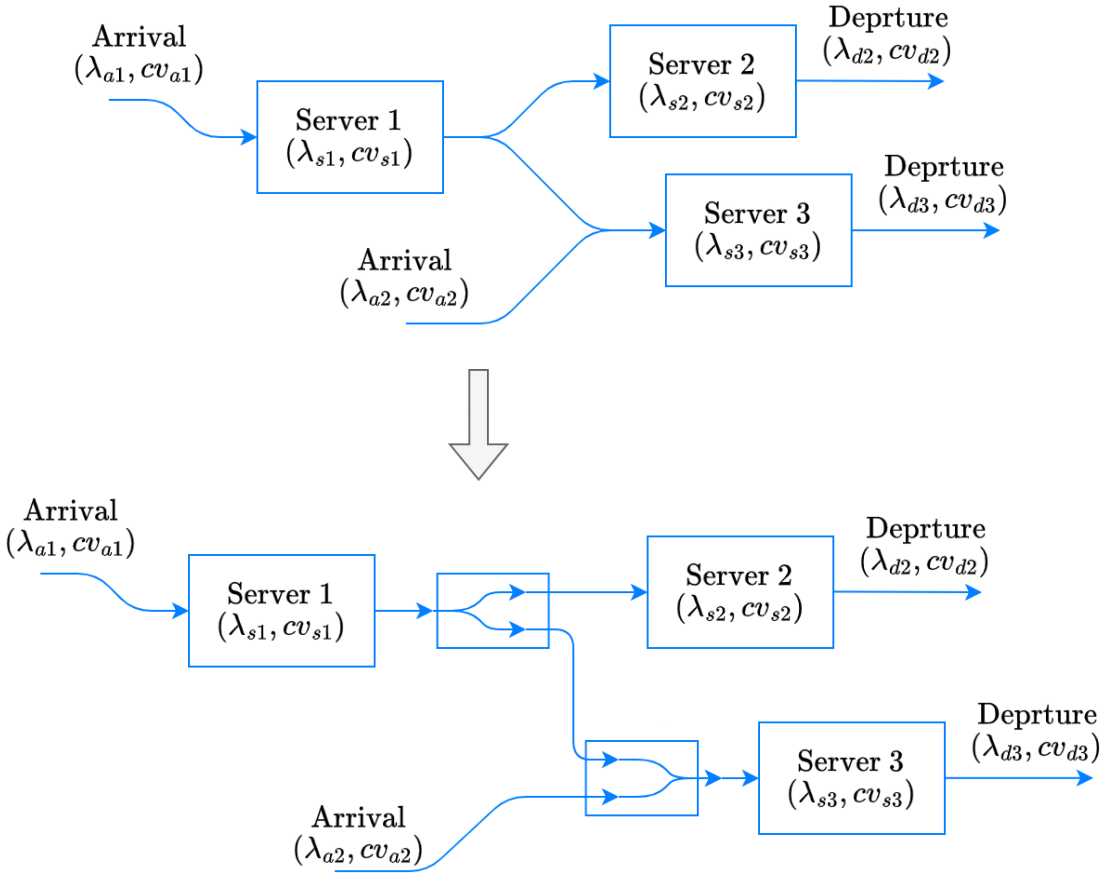
$$CT_q = \frac{(cv_a^2 + cv_s^2)}{2} \frac{u}{1-u} \frac{1}{\lambda_s}$$

$$cv_d^2 = u^2 cv_s^2 + (1-u^2) cv_a^2$$

  - Split

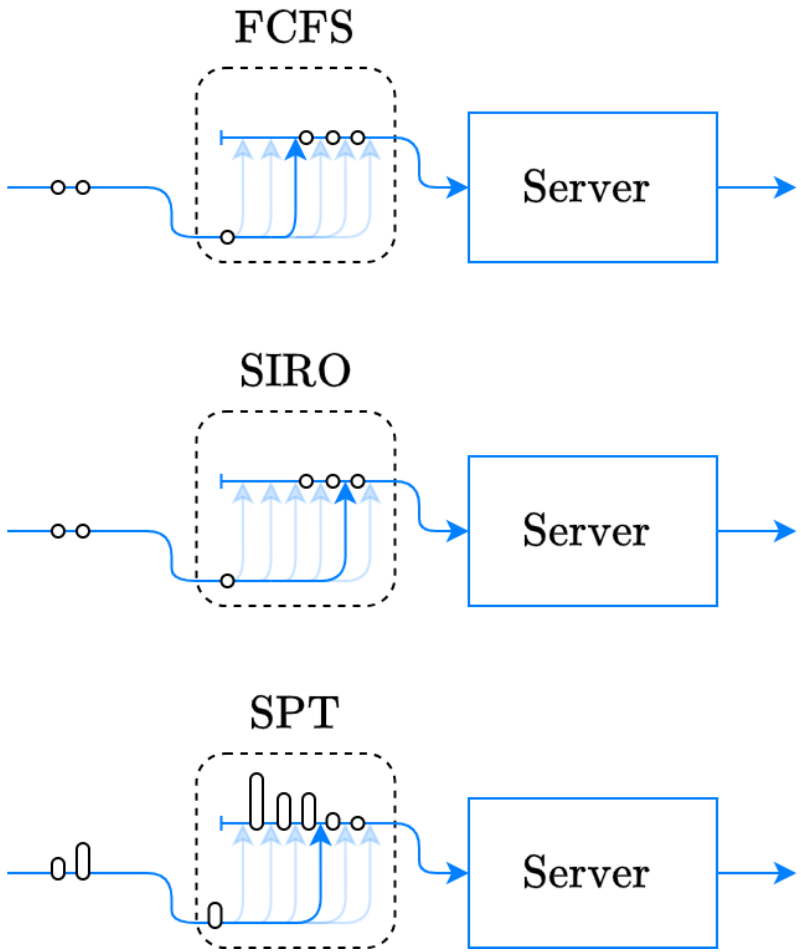
$$cv_d^2 = p\, cv_a^2 + (1-p)$$

  - Merge
    - The explicit formulas are too long to fit here



Queueing Network Analysis

- Drawbacks of decomposition with analytical models
  - Real world systems have correlated processes and the i.i.d. assumption is restrictive
  - The first come first serve (FCFS) sequencing rule is restrictive
    - Service in random order (SIRO) can better model sequencing based on unobservable features
    - Shortest processing time first (SPT) gives better cycle times
- Solution?
  - Machine learning



Machine Learning for Queueing Network Analysis

- How can machine learning be used for analyzing networks of queues?
  - Analysis of each node
  - Simulation can be used for generating training data
  - Pre-trained machine learning models
- Curse of dimensionality
  - High performance clusters
  - Sampling

Arrival process (a)		Service times (s)		Cycle time	Departure process (d)	
λ_a	cv_a	λ_s	cv_s	$E(CT)$	λ_d	cv_d
0.5	1	1	1	?	?	?

$$f_{E(CT)}(\lambda_a, cv_a, \lambda_s, cv_s) = E(CT)$$
$$f_{\lambda_d}(\lambda_a, cv_a, \lambda_s, cv_s) = \lambda_d$$
$$f_{cv_d}(\lambda_a, cv_a, \lambda_s, cv_s) = cv_d$$

Machine Learning for Queueing Network Analysis

- How can machine learning be used for analyzing networks of queues?
  - Analysis of each node
  - Simulation can be used for generating training data
  - Pre-trained machine learning models
- Curse of dimensionality
  - High performance clusters
  - Sampling

Arrival process (a)		Service times (s)		Cycle time	Departure process (d)	
λ_a	cv_a	λ_s	cv_s	$E(CT)$	λ_d	cv_d
0.5	1	1	1	?	?	?
0.2	1	2	1	0.55	0.2	1
0.3	1	1	1	1.42	0.3	1
0.1	1	2	1	0.52	0.1	1

$$f_{E(CT)}(\lambda_a, cv_a, \lambda_s, cv_s) = E(CT)$$
$$f_{\lambda_d}(\lambda_a, cv_a, \lambda_s, cv_s) = \lambda_d$$
$$f_{cv_d}(\lambda_a, cv_a, \lambda_s, cv_s) = cv_d$$

Machine Learning for Queueing Network Analysis

- How can machine learning be used for analyzing networks of queues?
  - Analysis of each node
  - Simulation can be used for generating training data
  - Pre-trained machine learning models
- Curse of dimensionality
  - High performance clusters
  - Sampling

Arrival process (a)		Service times (s)		Cycle time	Departure process (d)	
λ_a	cv_a	λ_s	cv_s	$E(CT)$	λ_d	cv_d
0.5	1	1	1	?	?	?
0.2	1	0.8	1	1.66	0.2	1
0.3	1	0.8	1	2	0.3	1
0.4	1	0.8	1	2.5	0.4	1
0.5	1	0.8	1	3.33	0.5	1
0.6	1	0.8	1	5	0.6	1
0.2	1	0.9	1	1.42	0.2	1
0.3	1	0.9	1	1.66	0.3	1
0.4	1	0.9	1	2	0.4	1
0.5	1	0.9	1	2.5	0.5	1
0.6	1	0.9	1	3.33	0.6	1
0.2	1	1	1	1.25	0.2	1
0.3	1	1	1	1.42	0.3	1
0.4	1	1	1	1.66	0.4	1
0.5	1	1	1	2	0.5	1
0.6	1	1	1	2.5	0.6	1

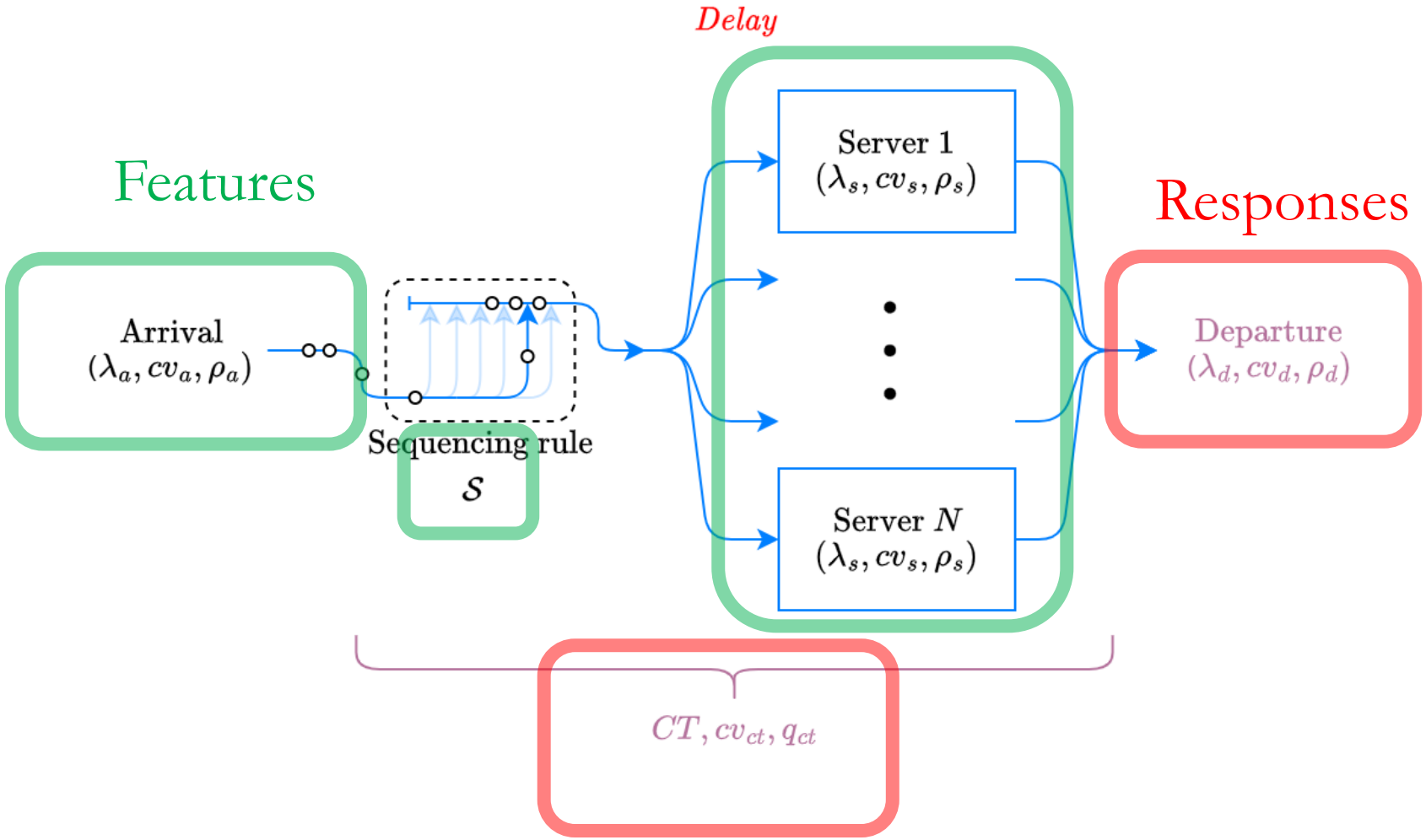
Building the Method Based on Supervised Learning

- Data requirement for the blocks with new parameters for first lag autocorrelation ρ_a and the number of servers N
  - For the delay block: $9 \times 14 \times 9 \times 1 \times 14 \times 9 \times 1 = 142884$ **~ 1 year**

Parameter	Range
λ_a	{0.1, 0.2,...,0.9}
cv_a	{0.1,...,1.4}
ρ_a	{−0.4,...,0.4}
λ_s	{1}
cv_s	{0.1,...,1.4}
ρ_s	{−0.4,...,0.4}
N	{1}

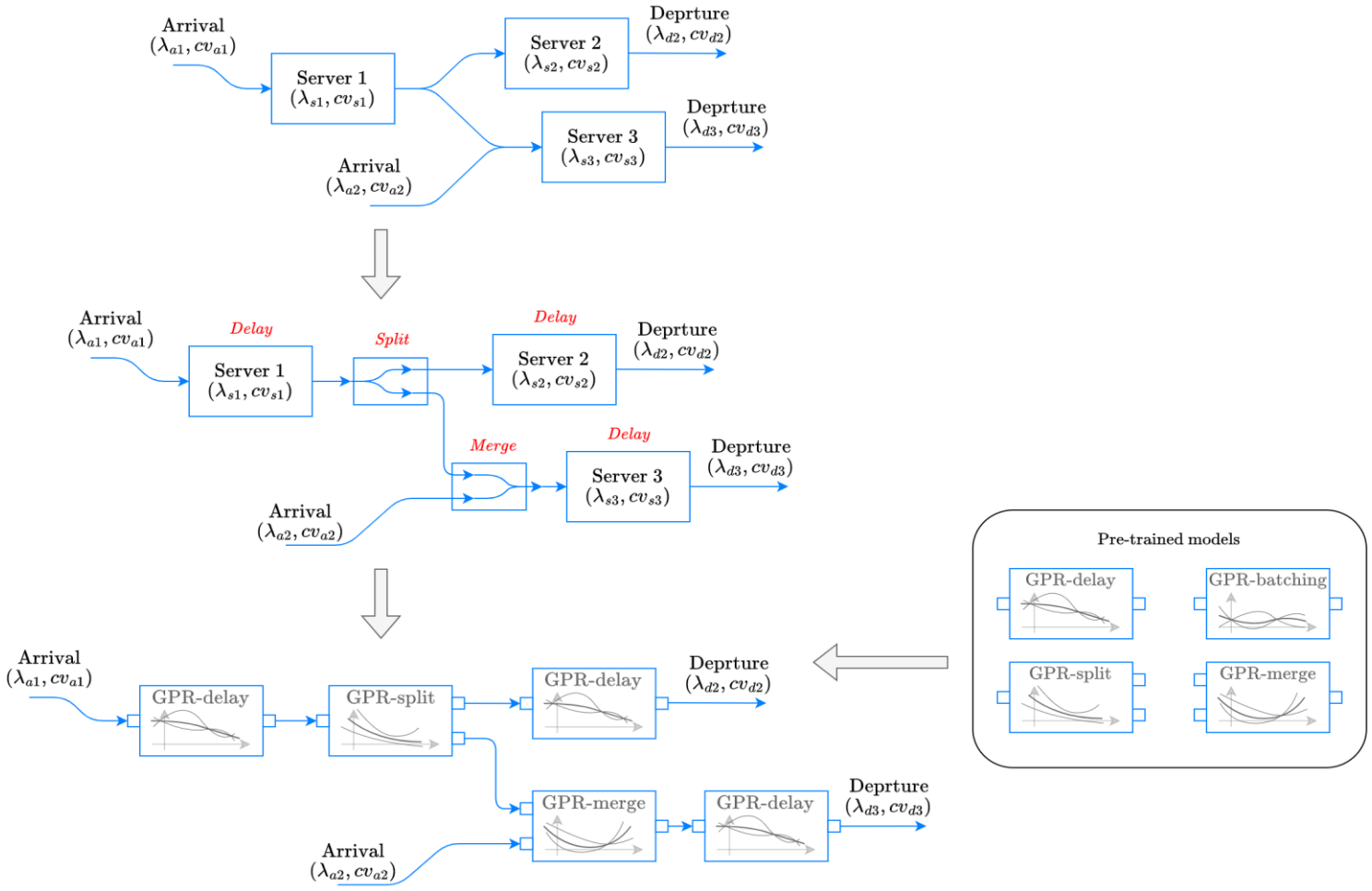
Building the Method Based on Supervised Learning

- The blocks used in decomposition
  - Delay
  - Split
  - Merge
  - Batching
- We use Gaussian process regression (**GPR**)



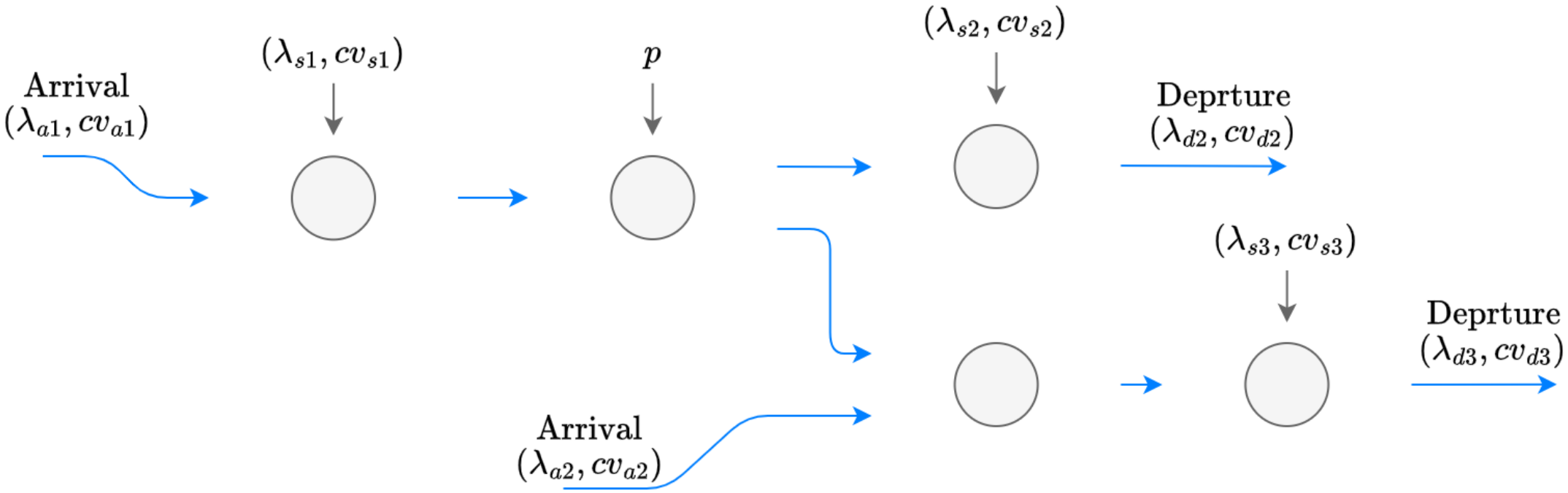
Building the Method Based on Supervised Learning

- The algorithm



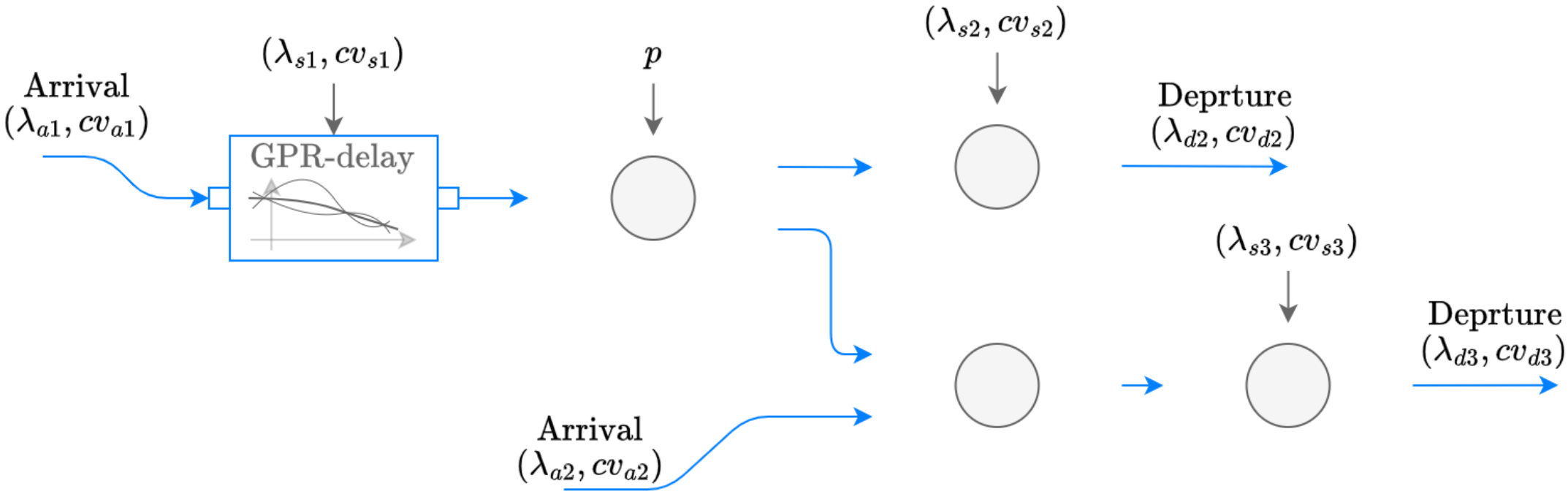
Building the Method Based on Supervised Learning

- The algorithm



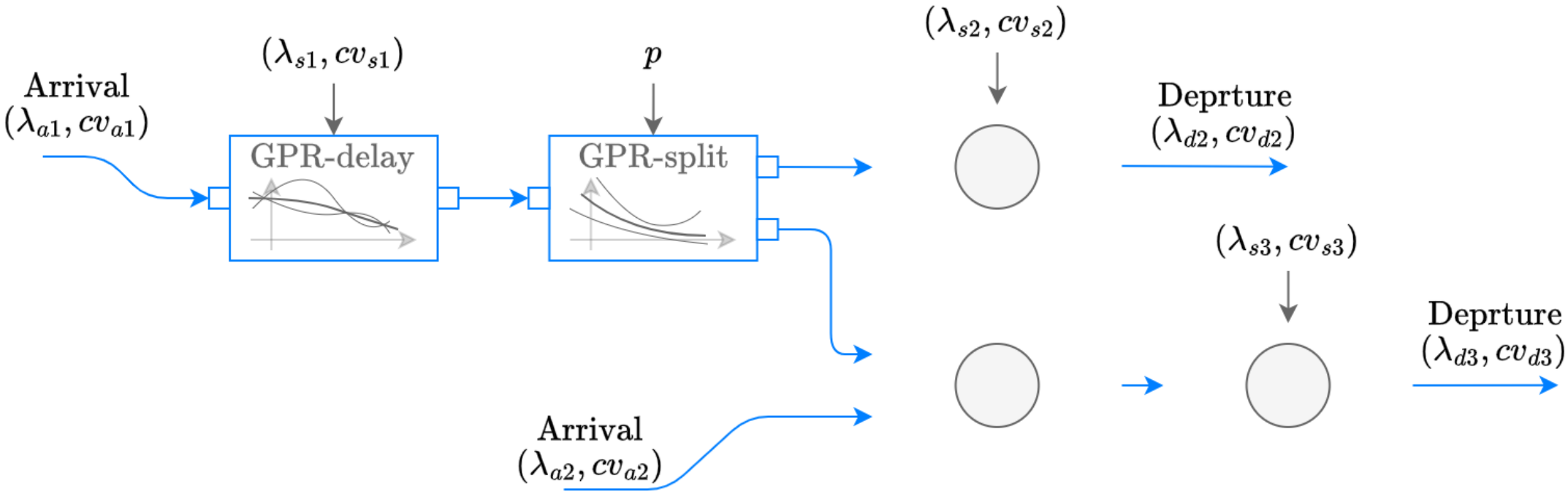
Building the Method Based on Supervised Learning

- The algorithm



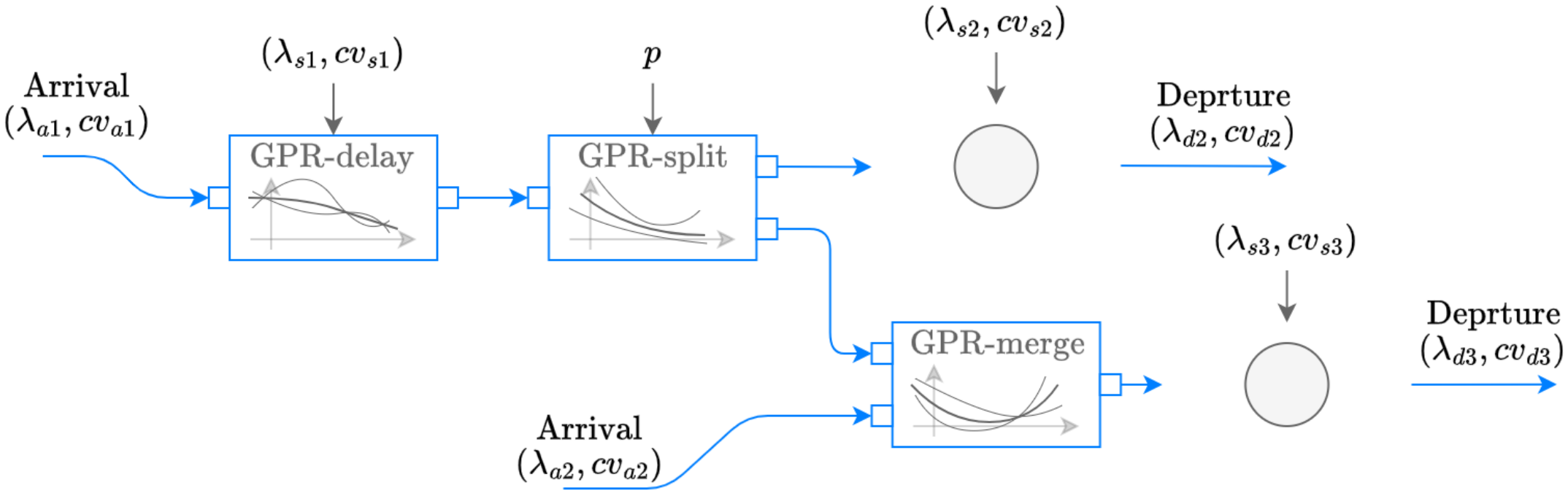
Building the Method Based on Supervised Learning

- The algorithm



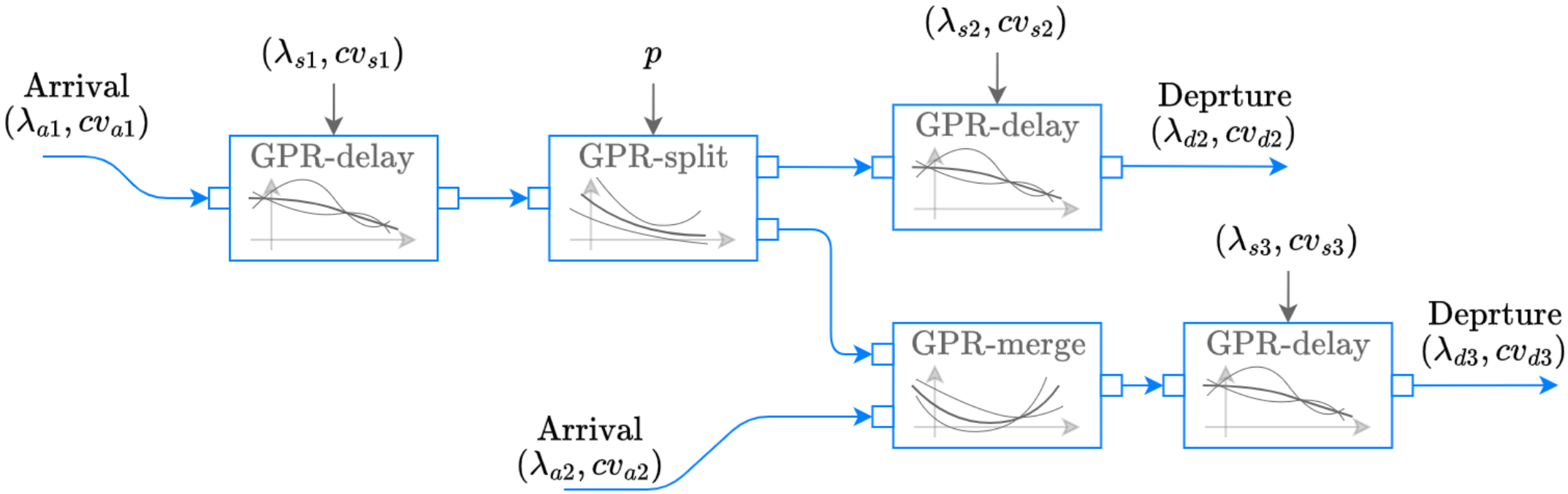
Building the Method Based on Supervised Learning

- The algorithm



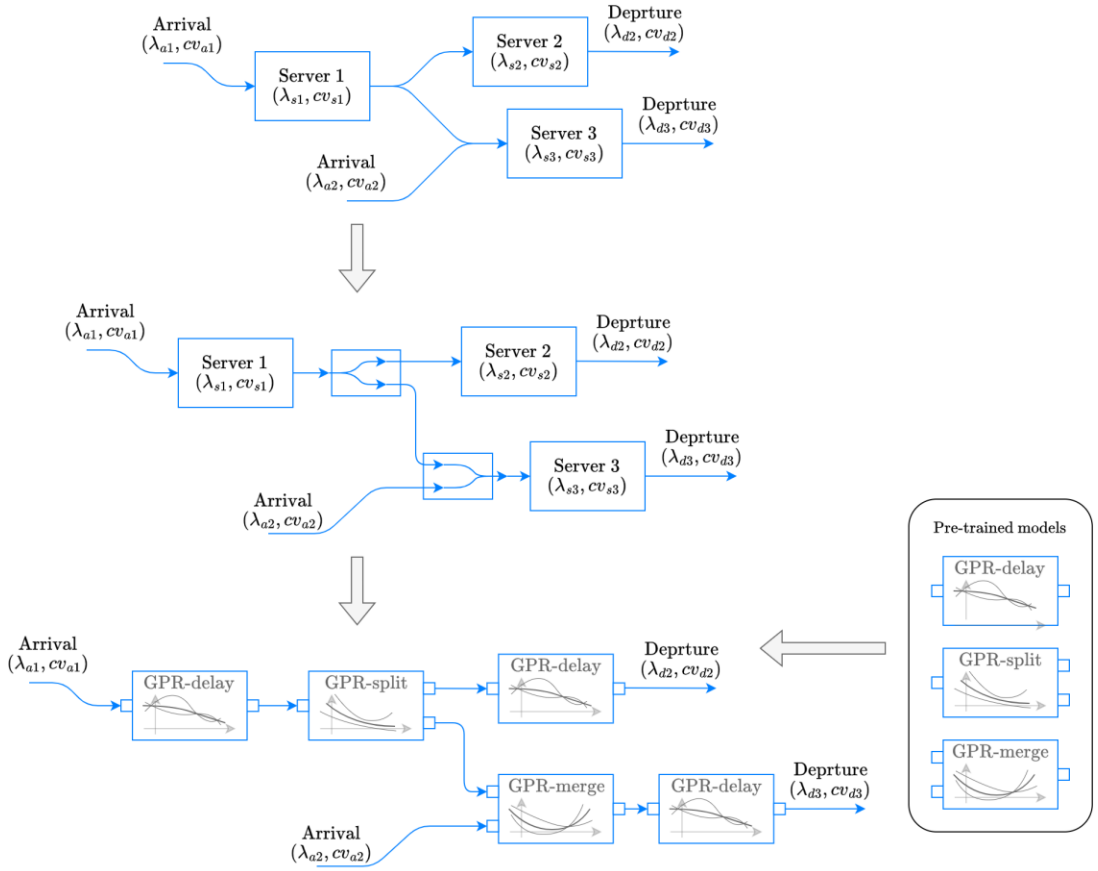
Building the Method Based on Supervised Learning

- The algorithm



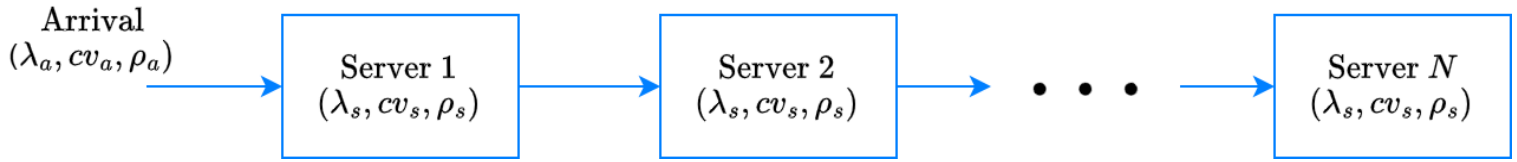
Building the Method Based on Supervised Learning

- How complex queueing systems can be analyzed?
  - Decomposition
- Drawbacks of decomposition with analytical models
  - Real world systems have correlated processes
  - The FCFS sequencing rule is restrictive
    - SIRO can better model sequencing based on unobservable features
    - SPT gives better cycle times
- Solution?
  - Machine learning
  - Pre-trained models → SLQNA (Supervised Learning based Queuing Network Analyzer)



Assessing the performance of SLQNA

- Serial lines



Parameter	Range
μ_a	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$
μ_s	$\{1\}$
cv_a, cv_s	$\{0.4, 0.6, 0.8, 1.0, 1.2\}$
ρ_a, ρ_s	$\{-0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4\}$
N	$\{5,10,15,20,25\}$

Range of parameters used for the serial line experiments with homogeneous stations

Assessing the performance of SLQNA

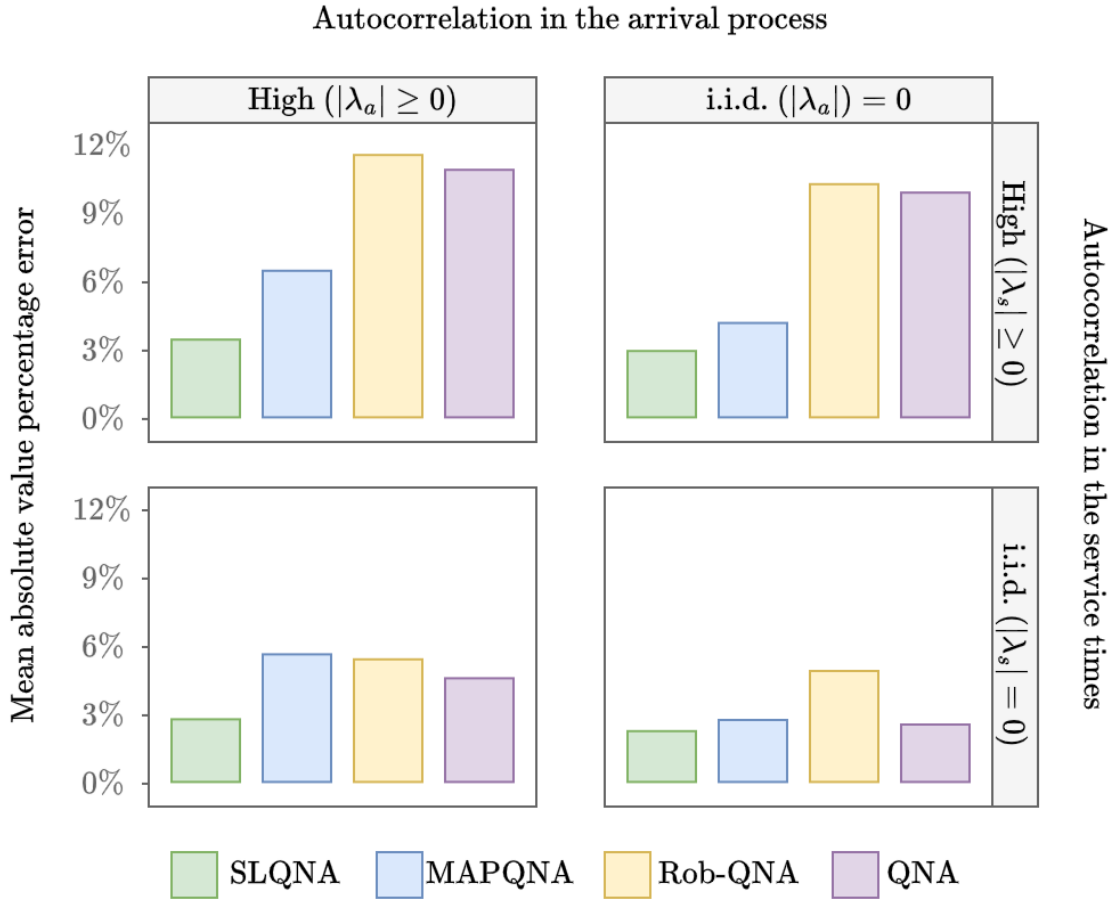
- Comparison with other methods (where it applies)
  - Analytical models (QNA) (Kuehn 1979)
    - Cannot consider autocorrelation
    - Assumes the FCFS discipline
  - MAPs (MAPQNA) (Horváth et al. 2010)
    - Parameter fitting for MAPs can be slow
    - The matrix geometric method can be slow
    - Assumes the FCFS discipline
  - Robust queuing (Rob-QNA) (Whitt and You 2020)
    - Cannot incorporate service time autocorrelation
    - Assumes the FCFS discipline

Assessing the performance of SLQNA

- Serial lines
  - SLQNA is fast, accurate, versatile

Mean absolute value percentage error:

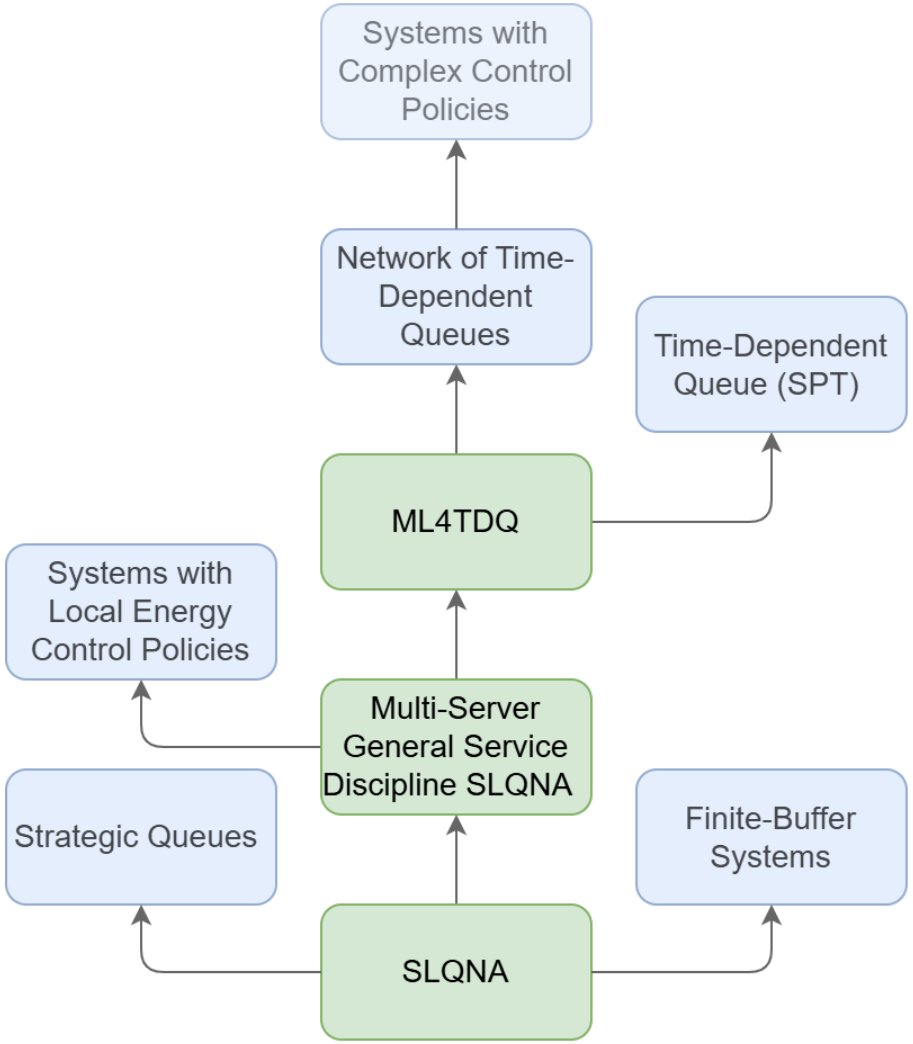
$$\text{MAPE} = \frac{|\boldsymbol{CT} - \widehat{\boldsymbol{CT}}|}{\widehat{\boldsymbol{CT}}} \times 100$$



Accuracy of different methods in predicting the cycle time for the production line experiments

Future Work

- Extensions in the time-dependent queueing area
  - Time-dependent queueing networks
  - Analysis of SPT time-dependent queues
- Extensions in stationary setting
  - Analysis of finite-buffer queueing networks
  - Extension to energy control
  - Strategic queues and MNL
- Systems with complex control policies



Thank you!

References

- Kuehn, Paul. “Approximate analysis of general queuing networks by decomposition.” IEEE Transactions on communications 27.1 (1979): 113-126.
- Whitt, Ward, and Wei You. "A robust queueing network analyzer based on indices of dispersion." Naval Research Logistics (NRL) 69.1 (2022): 36-56.
- Horváth, András, Gábor Horváth, and Miklós Telek. "A joint moments based analysis of networks of MAP/MAP/1 queues." Performance Evaluation 67.9 (2010): 759-778.

Appendix

Appendix

- Time Performance of SLQNA