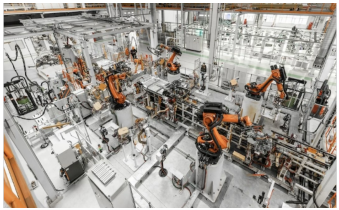
Diffusion-Based Policies for Dynamic Control of Stochastic Processing Networks

Jin Guang

Chicago Booth
joint work with Barış Ata, J. Michael Harrison, and Nian Si

2026 INFORMS Annual Meeting
November 2, 2026

Stochastic processing networks model many operational systems



Manufacturing

Sequencing jobs



Support centers

Alternative processing

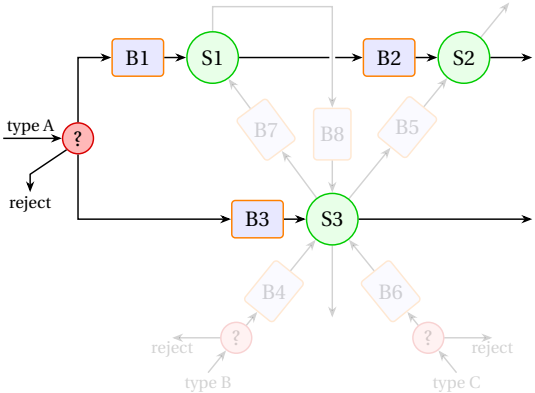


Data centers

Routing and admission control

- ✓ Random demand
- ✓ Congestion and delay
- ✓ Shared resources
- ✓ Dynamic decision making

Dynamic control for stochastic networks

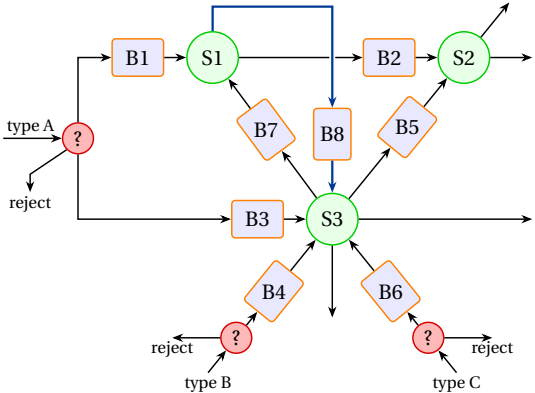


A three-station processing network (Harrison 1996)

A new type-A job arrives.

- Send it through S1 and S2.
- Send it directly to S3.
- Reject it and pay a penalty.

Dynamic control for stochastic networks



A three-station processing network (Harrison 1996)

Other jobs share the specialists.

Type B: S3 → S2.

Type C: S3 → S1 → S3.

Accepting more work can delay jobs throughout the network.

Objective

Minimize expected discounted or long-run average total cost.

Holding costs + rejection penalties.

Why network control is computationally difficult

The curse of dimensionality and high load

Poisson arrivals, exponential service: queue lengths form the MDP state.

M values per buffer

Typical scale: $M = O\left(\frac{1}{1-\rho}\right)$

ρ: utilization, $\rho \uparrow 1$.

Truncated MDP state space



Exponential growth in I.

Why network control is computationally difficult

The curse of dimensionality and high load

Poisson arrivals, exponential service: queue lengths form the MDP state.

M values per buffer

Typical scale: $M = O\left(\frac{1}{1-\rho}\right)$

ρ: utilization, $\rho \uparrow 1$.

Truncated MDP state space

M^I

I buffers

Exponential growth in I.

General interarrival and service times

Queue lengths alone are generally not Markov.

$$\text{Augmented state} = \Big(\underbrace{\text{queue lengths}}_{\text{discrete}}, \underbrace{\text{arrival and service ages}}_{\text{continuous}}\Big).$$

Related literature

Original network model

Approximately solve the MDP or learn policies directly.

Approximate policy iteration

Sauré, Patrick and Puterman (2015).

Reinforcement learning

Dai and Gluzman (2022).
Che, Dong and Namkoong (2024).

Fluid and diffusion approximations

Solve a fluid or diffusion model, then construct a network policy.

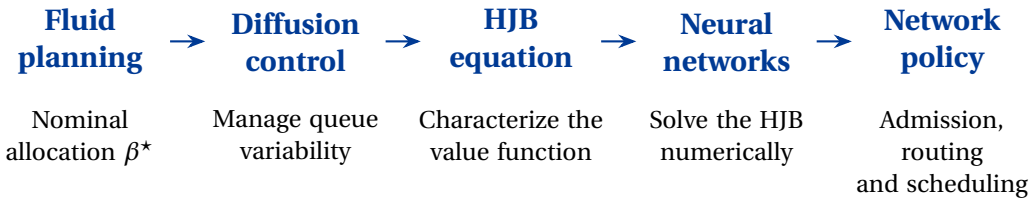
Fluid approximation

Avram, Bertsimas and Ricard (1995).
Maglaras (2000).
Bertsimas et al. (2015).

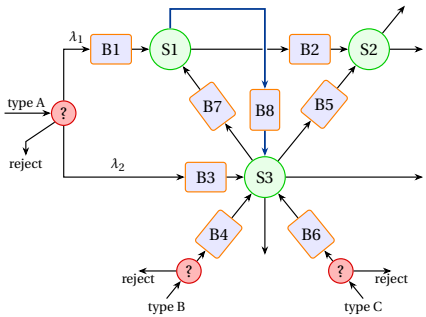
Diffusion approximation

Kumar and Muthuraman (2004).
Ata, Harrison and Si (2024, 2025).

Our approach



Fluid planning



$(\alpha_A, \alpha_B, \alpha_C) = (0.5, 0.25, 0.25)$,

$m = (2, 3, 1, 1, 1, 1, 2, 1)$.

Static planning problem

m_j: mean service time.

λ_1, λ_2: type-A route flows.

a_k: admitted rate; c_k: rejection penalty.

$$\underset{\rho,\lambda_1,\lambda_2,a\geq 0}{\text{lexmin}} \Big(\underbrace{c^\top(\alpha - a)}_{\text{1st priority}}, \quad \underbrace{\rho}_{\text{2nd priority}} \Big)$$

$$\begin{aligned}
\text{s.t.} \quad & \lambda_1 + \lambda_2 = a_A, \\
\text{S1:} \quad & m_1\lambda_1 + m_7 a_C \leq \rho, \\
\text{S2:} \quad & m_2\lambda_1 + m_5 a_B \leq \rho, \\
\text{S3:} \quad & m_3\lambda_2 + m_4 a_B + (m_6 + m_8) a_C \leq \rho, \\
& a \leq \alpha, \qquad \rho \leq 1.
\end{aligned}$$

$\lambda_1^\star = \lambda_2^\star = 0.25, \qquad \rho^\star = 1.$

Diffusion control

Adjust the nominal allocation to manage variability.

$Z(t)$: approximating queue vector in diffusion units.

$$dZ(t) = \underbrace{b(u(t))\,dt}_{\text{controlled net flow}} + \underbrace{dX(t)}_{\text{variability}} + \underbrace{H\,dL(t)}_{\text{boundary correction}} .$$

$u(t) \in \Theta$: state-dependent adjustments around $\beta^\star$.

H is fixed; L_i increases only when buffer i is empty.

Brownian control problem (BCP)

Minimize discounted or long-run average holding and rejection costs.

Solving the HJB equation

The equation characterizes the value and its marginal costs.

$$\mathscr{L}\widetilde{V}(z) + \min_{u\in\Theta}\{\ell(z,u) + b(u)\cdot\nabla\widetilde{V}(z)\} = \gamma\widetilde{V}(z) + \xi, \qquad z > 0,$$

$$(H^{\mathsf{T}}\nabla\widetilde{V}(z))_i = 0 \quad \text{on } z_i = 0, \qquad i = 1,\dots,m.$$

$\ell(z,u)$: holding and rejection cost rate.

$\mathscr{L} = \frac{1}{2}\Gamma : \nabla^2$ captures variability.

Discounted: $\xi = 0$. Long-run average: $\gamma = 0$, with optimal cost rate ξ.

Neural network solution via deep BSDE methods

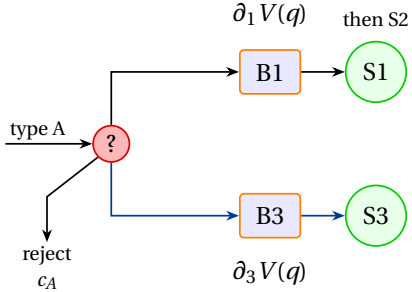
Approximate $\widetilde{V}$ and $\nabla\widetilde{V}$; use the gradient to construct the policy.

E, Han and Jentzen (2017); Ata, Harrison and Si (2025).

Diffusion-based policy: admission and routing

$\partial_i V(q)$: marginal cost of additional work in buffer i.

A type-A job arrives. Compare three costs.



Cost per arriving job

Admit to B1: $\partial_1 V(q)$,

Admit to B3: $\partial_3 V(q)$,

Reject: c_A.

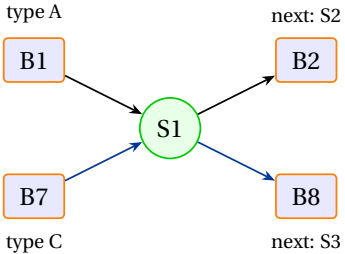
Choose the smallest of $\partial_1 V(q)$, $\partial_3 V(q)$, and c_A.

The marginal admission costs include future congestion across the network.

Diffusion-based policy: scheduling

$\mu_j = 1/m_j$: service rate of activity j.

S1 can serve B1 or B7. Which choice is better?



Marginal cost change per unit time

Serve B1: $\mu_1\left[\partial_2 V(q) - \partial_1 V(q)\right]$,

Serve B7: $\mu_7\left[\partial_8 V(q) - \partial_7 V(q)\right]$,

Idle: 0.

Choose the smallest cost change. Exclude empty buffers.

Each marginal value depends on the whole network state.

Three-station control at critical load

Arrivals and service	Poisson $\alpha = (0.5, 0.25, 0.25)$; exponential $\mu = (\frac{1}{2}, \frac{1}{3}, 1, 1, 1, 1, \frac{1}{2}, 1)$
Nominal station loads	$(\rho_1, \rho_2, \rho_3) = (1, 1, 1)$ (all admitted; A split 50/50)
Costs	Holding $h = (6, 3, 6, 6, 1, 12, 7, 6)$; rejection per job $c = (396, 480, 1080)$

Discounted cost

Total cost; $\delta = 0.01$.

Policy	Cost
MDP (cap 44)	8081.23 ± 10.07
Ours	8134.52 ± 10.11
Gap	+0.66%

Long-run average cost

Cost per unit time.

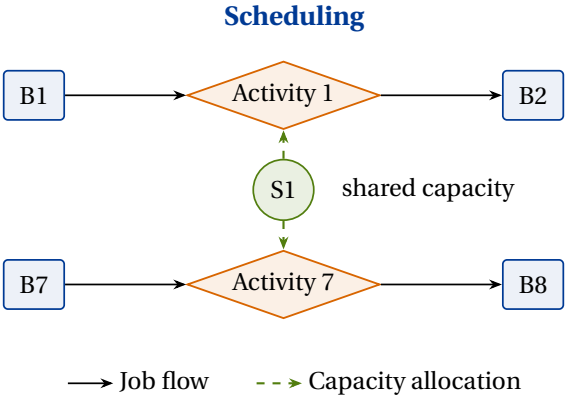
Policy	Cost rate
MDP (cap 44)	108.65 ± 0.09
Ours	109.14 ± 0.09
Gap	+0.44%

Exact aggregation: merge B3 and B8 into $q_3 + q_8$ (S3, $\mu = 1$, $h = 6$, exit).

MDP states: 752,538,150 $\longrightarrow$ 115,775,100 $\quad (\sum_i q_i \le 44$; 6.5-fold reduction).

$\pm$: Monte Carlo SE. Gap: percentage above the MDP cost.

Stochastic processing network (SPN)



Control selects activities subject to resource capacities.

Buffers

Jobs awaiting processing.

Activities

Ways to process or admit jobs.

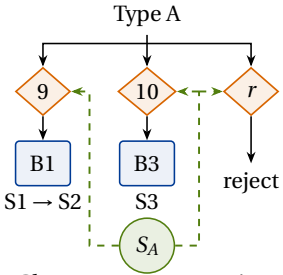
Resources

Capacity shared by activities.

Stochastic processing network (SPN)

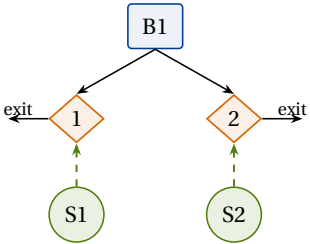
Buffers hold jobs; activities use shared resource capacity.

Alternative routing



Choose a route or reject.

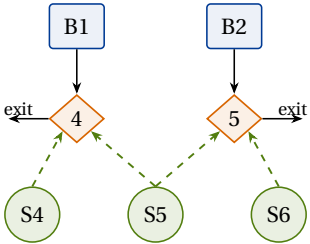
Alternative processing



Choose a processing activity.

Pesic and Williams (2016).

Multi-resource processing



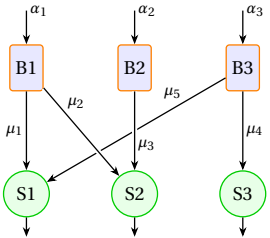
Activities compete for S5.

Dai and Lin (2005).

Control selects activities subject to resource capacities.

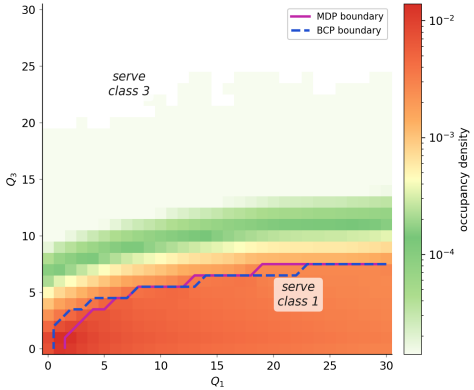
Pesic–Williams: policy and performance

Holding costs: $h = (1, 2, 3)$. Nominal station loads: $(\rho_1, \rho_2, \rho_3) \approx (0.967, 0.967, 0.95)$.



Pesic and Williams (2016)

Policy	Discounted	Average
MDP	2581.50 ± 3.05	58.94 ± 0.23
Ours	2586.55 ± 3.04	59.73 ± 0.23
Gap	+0.20%	+1.34%



MDP Ours $Q_2 = 0$.

General distributions: Pesic–Williams

Long-run average cost; fixed mean arrival and service times.

CoV	MDP (Exp.)	Ours (Exp.)	Ours (CoV)	Change
0.5	18.41 ± 0.03	18.74 ± 0.03	18.15 ± 0.03	**−3.13%**
2	—	228.55 ± 1.95	221.21 ± 1.71	**−3.21%**

CoV: coefficient of variation.

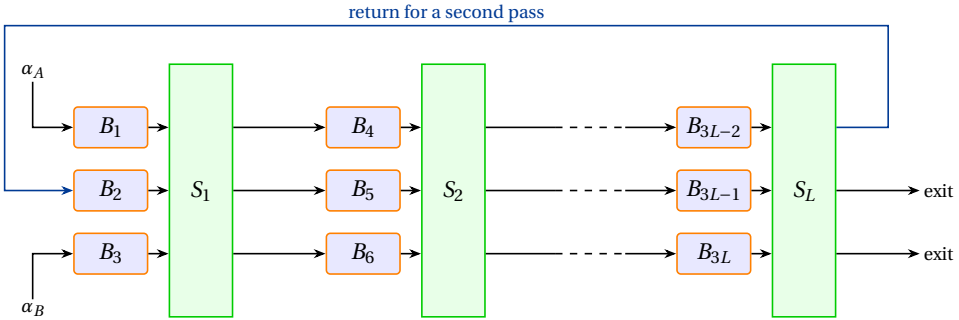
Gamma: CoV 0.5. Lognormal: CoV 2.

Account for variability through the diffusion covariance.

Change: Ours (CoV) relative to Ours (Exp.). ±: Monte Carlo SE.
At CoV 2, the MDP grid misses 11.7% of state lookups.

Scalable control: Extended six-class network

L stations, $3L$ buffers: increase the state dimension from 6 to 21.



Each station serves three buffers; one stream returns for a second pass.

$\alpha_A = \alpha_B = 9/140, \quad \rho_s = 0.9$ for every station, $\quad h_i = 1.$

Scalable control: Extended six-class network

Long-run average number of jobs.

Stations	Buffers	Max-pressure Dai–Lin (2005)	LBFS	PPO Dai–Gluzman (2022)	Best LLM Baek (2026)	Ours This work
2	6	16.66	14.36	14.13	13.87	13.67
3	9	24.75	23.47	23.27	21.27	21.27
4	12	32.41	31.62	32.17	29.37	27.96
5	15	40.03	41.93	39.30	36.84	35.85
6	18	47.68	50.24	51.47	42.95	43.88
7	21	54.18	60.61	55.12	55.07	51.15

Best LLM: `gpt-5.6-sol`, best reported variant for each instance.

Thank you!

Questions & Answers

Appendix: policy for a general network

Use marginal values to compare feasible activities.

R^j: mean net depletion from activity j.

$$\text{Marginal cost rate for activity } j: \qquad -(R^j)^\top \nabla V(q) - c^\top K^j.$$

$$x(q) \in \arg\min_{x \in \mathcal{X}(q)} \sum_j \left[-(R^j)^\top \nabla V(q) - c^\top K^j\right] x_j.$$

$\mathcal{X}(q)$ imposes resource capacities and buffer availability.

Scheduling at S1

$$\mu_1[\partial_2 V(q) - \partial_1 V(q)],$$

$$\mu_7[\partial_8 V(q) - \partial_7 V(q)], \quad 0.$$

Type-A admission and routing

$$\partial_1 V(q), \quad \partial_3 V(q), \quad c_A.$$

Compare costs per arriving job.

Minimize the cost change per unit time.
The discounted and average-cost objectives require separate value functions.

Appendix: from fluid planning to diffusion control

First order: the nominal allocation $\beta^\star$.

$$R\beta^\star = \alpha^\star, \qquad A\beta^\star = e.$$

$\alpha^\star$: critical external buffer input rates; input activities are included in R.

Second order: deviations from the fluid allocation.

$$Z^{(r)}(t) = \frac{Q(rt)}{\sqrt{r}}, \qquad Y^{(r)}(t) = \frac{r\beta^\star t - T(rt)}{\sqrt{r}}.$$

$$Z^{(r)}(t) \approx z + X(t) + \zeta t + RY^{(r)}(t), \qquad \zeta = \sqrt{r}(\alpha - R\beta^\star).$$

Modified BCP: $dY = u\,dt + D\,dL, \quad H = RD.$

$$dZ = (\zeta + Ru)\,dt + dX + H\,dL.$$

Bounded u and fixed H are computational model choices.

Appendix: Brownian control problem

$\ell(z,u)$: total cost rate (holding and rejection).

$$\inf_{u(\cdot)} \begin{cases} \mathbb{E}_z\left[\int_0^\infty e^{-\gamma t}\ell(Z(t),u(t))\,dt\right], & \text{discounted,} \\ \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_z\left[\int_0^T \ell(Z(t),u(t))\,dt\right], & \text{long-run average.} \end{cases}$$

$$\begin{aligned} \text{s.t.} \quad & dZ(t) = b(u(t))\,dt + dX(t) + H\,dL(t), \\ & Z(0) = z, \quad Z(t) \geq 0, \quad u(t) \in \Theta, \\ & L_i(0) = 0, \quad dL_i(t) \geq 0, \quad Z_i(t)\,dL_i(t) = 0. \end{aligned}$$

$\widetilde{V}$: (relative) value; ξ: optimal average cost rate.

In network units, write this value as $V(q)$.

Appendix: HJB, reflection, and the training identity

$$\mathscr{L}\widetilde{V} + \min_{u\in\Theta}\{\ell(z,u) + b(u)\cdot\nabla\widetilde{V}\} = \gamma\widetilde{V} + \xi, \qquad \mathscr{L} = \tfrac{1}{2}\operatorname{tr}(\Gamma\nabla^2).$$

$$(H^i)^\top \nabla\widetilde{V}(z) = 0 \quad \text{on } z_i = 0.$$

Simulate a reference diffusion $d\bar{Z} = b_0\,dt + dX + H\,dL$.

$$F(z,g) = \min_{u\in\Theta}\{\ell(z,u) + [b(u) - b_0]\cdot g\}.$$

$$0 = e^{-\gamma T}\widetilde{V}(\bar{Z}_T) - \widetilde{V}(\bar{Z}_0) - \int_0^T e^{-\gamma t}\nabla\widetilde{V}(\bar{Z}_t)\cdot dX_t$$
$$+ \int_0^T e^{-\gamma t}\big[F(\bar{Z}_t, \nabla\widetilde{V}(\bar{Z}_t)) - \xi\big]\,dt.$$

Discounted: $\xi = 0$. Long-run average: $\gamma = 0$.

Appendix: learning the value and its gradient

Approximate $\widetilde{V}$ and $\nabla\widetilde{V}$ with two neural networks.

Discounted objective

Set $\xi = 0$ in the stochastic identity.

$$\min_{\eta,\phi} \mathbb{E}[\mathscr{R}_T(\widetilde{V}_\eta, G_\phi)^2].$$

Fit the identity along simulated diffusion paths.

Long-run average objective

Set $\gamma = 0$ and write $\mathscr{B}_T = \xi T$.

$$\min_{\eta,\phi} \mathrm{Var}(\mathscr{B}_T/T), \qquad \widehat{\xi} = \mathbb{E}[\mathscr{B}_T]/T.$$

A constant shift in the relative value is irrelevant.

Return to network units

$$V(q) = r^{3/2}\widetilde{V}(q/\sqrt{r}), \qquad \partial_i V(q) \approx r\,[G_\phi(q/\sqrt{r})]_i.$$

The reported average-cost runs use occupancy restarts and a refreshed target network.

Appendix: what the cost comparisons measure

Evaluate the implemented policy in the original queueing network.

	Discounted	Long-run average
Criterion	Total discounted cost, $\delta = 0.01$	Cost per unit time
Training seeds	One	Three for each small network
Evaluation paths	100,000	1,000
Initial transient	Empty start	Warm-up 10,000
Scored horizon	1,200	50,000 after warm-up

Interpret the uncertainty and the comparator.

- $\pm$ measures Monte Carlo error conditional on the trained policies.
- Average-cost cells for our policy average the three fixed policies before computing the SE.
- The MDP reference is an implemented truncated-state policy.

Appendix: making the three-station MDP tractable

Aggregate B3 and B8 exactly under exponential service.

Both use S3, have $\mu = 1$ and $h = 6$, and leave after service.

$$q \longmapsto (q_1, q_2, q_3 + q_8, q_4, q_5, q_6, q_7).$$

For total queue length at most 44:

$$\binom{52}{8} = 752{,}538{,}150 \quad \longrightarrow \quad \binom{51}{7} = 115{,}775{,}100 \text{ states.}$$

Truncation sensitivity of the average-cost Bellman value

Total-queue cap	24	40	44
Bellman average cost	108.660	108.6375	108.637

Aggregation is exact for this model; state-space truncation remains an approximation.

Appendix: general distributions and model accuracy

Change variability while holding the mean rates fixed.

CoV	Arrival and service distributions	Diffusion input
0.5	Gamma (Erlang-4)	Updated covariance Γ
1	Exponential	Baseline covariance
2	Lognormal	Updated covariance Γ

What is approximated?

The diffusion covariance uses primitive first and second moments.

This does not imply that exact network costs depend only on those moments.

What is the benchmark?

The exponential MDP policy is reused under the new distributions.

At CoV 2, its state grid is inadequate for Criss-cross and Pesic–Williams.

Appendix: Extended six-class network parameters

Every station serves three classes.

$$(m_{3s-2}, m_{3s-1}, m_{3s}) = \begin{cases} (8,2,4), & s \text{ odd}, \\ (6,7,1), & s \text{ even}. \end{cases}$$

$$\alpha_A = \alpha_B = 9/140, \qquad \rho_s = (9/140)\,(m_{3s-2} + m_{3s-1} + m_{3s}) = 0.9.$$

Evaluation protocols

Comparison	Source and evaluation
Ours / MP / LBFS / Best LLM	Baek (2026) simulator; 20 runs, 2×10^6 scored events per run.
PPO	Dai and Gluzman (2022), Table 4; their published simulation protocol.
Our policy training	One seed per network; 128k updates using diffusion paths.

Appendix: computation and comparison scope

Reported offline computation; hardware and tasks differ.

Stations	Ours: one H100	PPO: 96 CPU cores
2	47 min	100 min
3	54 min	146 min
4	77 min	202 min
5	82 min	424 min
6	107 min	862 min
7	116 min	1522 min

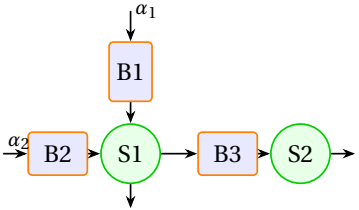
Best LLM: `gpt-5.6-sol` (Baek, 2026).

L1: 3,100–3,600 s of sandbox computation per instance; inference time unreported.
This does not cover the selected L2 variants or define a comparable total design time.

Ours: 128k diffusion-training updates per network. No hardware speedup claim.

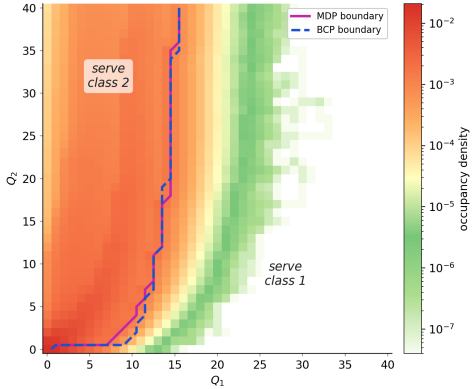
Appendix: Criss-cross policy and performance

Holding costs: $h = (1.5, 1, 1)$. Station loads: $(\rho_1, \rho_2) = (0.975, 0.95)$.



Harrison and Wein (1989)

Policy	Discounted	Average
MDP	1682.02 ± 2.06	56.79 ± 0.39
Ours	1685.97 ± 2.06	57.18 ± 0.38
Gap	+0.23%	+0.68%



MDP Ours $Q_3 = 0$.

Appendix: Criss-cross with general distributions

Long-run average cost; fixed mean arrival and service times.

CoV	MDP (Exp.)	Ours (Exp.)	Ours (CoV)	Change
0.5	16.82 ± 0.05	17.53 ± 0.05	16.33 ± 0.05	**−6.83%**
2	—	224.68 ± 2.69	222.39 ± 2.78	**−1.02%**

CoV: coefficient of variation.

Gamma: CoV 0.5. Lognormal: CoV 2.

Account for variability through the diffusion covariance.

Change: Ours (CoV) relative to Ours (Exp.). ±: Monte Carlo SE.
At CoV 2, the MDP grid misses 26.3% of state lookups.

Appendix: three-station general distributions

Fixed mean arrival and service times; vary their coefficient of variation (CoV).

Original rejection costs

Policy	CoV 0.5	CoV 2
MDP (Exp.)	73.15 ± 0.06	163.03 ± 0.16
Ours (Exp.)	73.94 ± 0.06	162.70 ± 0.16
Ours (CoV)	74.31 ± 0.04	170.25 ± 0.19
Change	**+0.49%**	**+4.64%**

Rejection costs $\times 4$

Policy	CoV 0.5	CoV 2
MDP (Exp.)	133.57 ± 0.25	372.33 ± 0.62
Ours (Exp.)	136.65 ± 0.25	380.48 ± 6.67
Ours (CoV)	121.30 ± 0.16	367.53 ± 0.73
Change	**−11.23%**	**−3.40%**

The effect of variability depends on admission costs.

Exp.: exponential training. CoV: retrained using the displayed variability.
Change: Ours (CoV) relative to Ours (Exp.). ±: Monte Carlo SE.