Dispatching to Fluid Queues

Esa Hyytiä[1,2] Runhan Xie[3] Rhonda Righter[3]

University of Iceland[1]
Aalto University[2]
University of California Berkeley[3]

INFORMS APS – Georgia Tech, Atlanta, United States
June 30, 2025

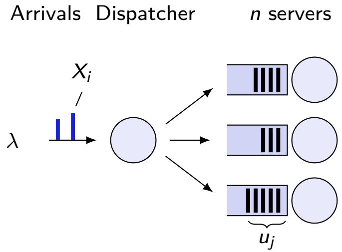
Outline:

1. Dispatching Problem
2. Fluid Dispatcing Problem
3. Theoretical Results
4. Numerical Examples

Dispatching Problem

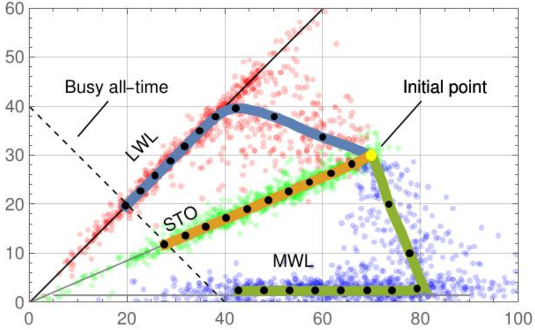
Control problem

1. n parallel FCFS servers
2. Job dispatching upon arrival
3. Poisson arrival process with rate λ
4. i.i.d. job sizes $X_i \sim X$
5. Jobsize X_i and backlogs u_j are known (upon arrival)
6. Minimize mean waiting time



Fundamentally a Markov Decision Problem
. . . in n-dimensional continuous space
. . . analytically intractable (when **both** jobsize and backlogs are known)

Sample Realizations with Heuristics



LWL Least-Work-Left chooses the queue with shortest backlog
(load balancing)

STO Straight-to-the-Origin tries to maintain ratio $u_1 : u_2$ constant
(fixed ratio on backlogs)

MWL Most-Work-Left chooses the queue with longest backlog
(load unbalancing, except when $u_2 < 3$ here)

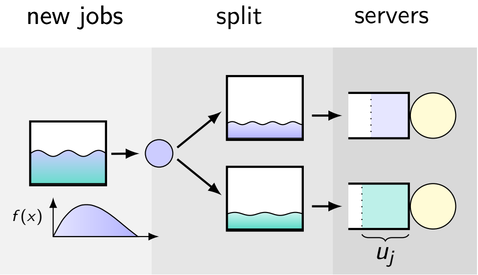
1) These policies ignore the size of the new job

Fluid approximation does well!

Routing Fluid to Parallel Servers

Fluid Control Problem

1. Fluid arrives at rate λ
2. Fluid consists of particles with size density $f(x)$ with $\mathbb{E}[X] = 1$
3. Dispatching is based on i) particle sizes and ii) server backlogs u_j
4. n parallel FCFS servers with service rates $1/n$



i) **Stability:** $\rho < 1$ $\quad (\rho = \lambda\,\mathbb{E}[X] = \lambda$ as $\mathbb{E}[X] = 1)$

ii) **Objective:** Minimize mean waiting time

Dispatching problem without stochastic fluctuations!

Fluid System Dynamics

Control action $\alpha(\mathbf{u})$ defines

1. Server-specific loads ρ_j (at state $\mathbf{u}$)
2. Server-specific rates λ_j

$$\sum_j \rho_j = \rho \text{ and } \sum_j \lambda_j = \lambda$$

Control $\alpha(\mathbf{u})$ thus defines *drainage rates*

$$\boxed{\dot{u}_j(t) = \frac{d}{dt} u_j(t) = \frac{1}{n} - \rho_j(t)}$$

and state-dependent *cost rates*

$$\boxed{\dot{c}_j(t) = \lambda_j(t) \cdot u_j(t)}$$

Optimization problem

Determine control $\alpha(\mathbf{u})$ to minimize total cost (=value function) for a given initial state $\mathbf{u}$

$$\sum_j \left(\int_0^T \dot{c}_j(t)\, dt \right) =: \min$$

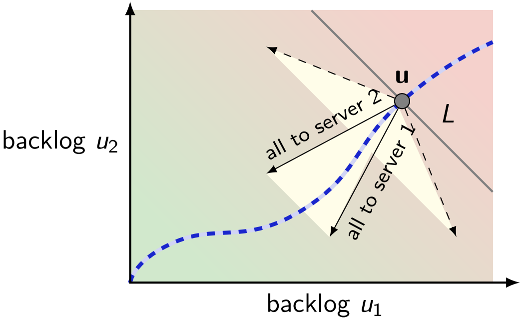
With *any* work-conserving policy ($\rho < 1$) the system empties at time

$$T = \frac{u_1 + \ldots + u_n}{1 - \rho}$$

Paths and Control

- Control $\alpha(\mathbf{u})$ defines the server-specific loads ρ_j
- The server-specific loads ρ_j define the trajectory how backlogs are emptied
- Not all paths are feasible
- Some backlogs may also increase!

ρ_j define the path	$\dot{u}_j = \frac{1}{n} - \rho_j$
λ_j and u_j define costs	$\dot{c}_j = \lambda_j \cdot u_j$



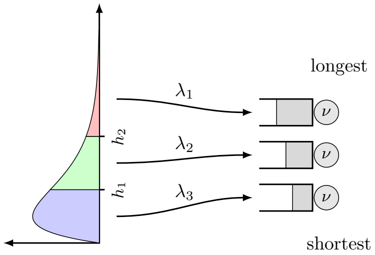
Problem: Find the optimal path to the origin!

(but the ρ_j can be realized many ways, each yielding different cost rate!)

Structural Results: Short to Short (for short)

Theorem

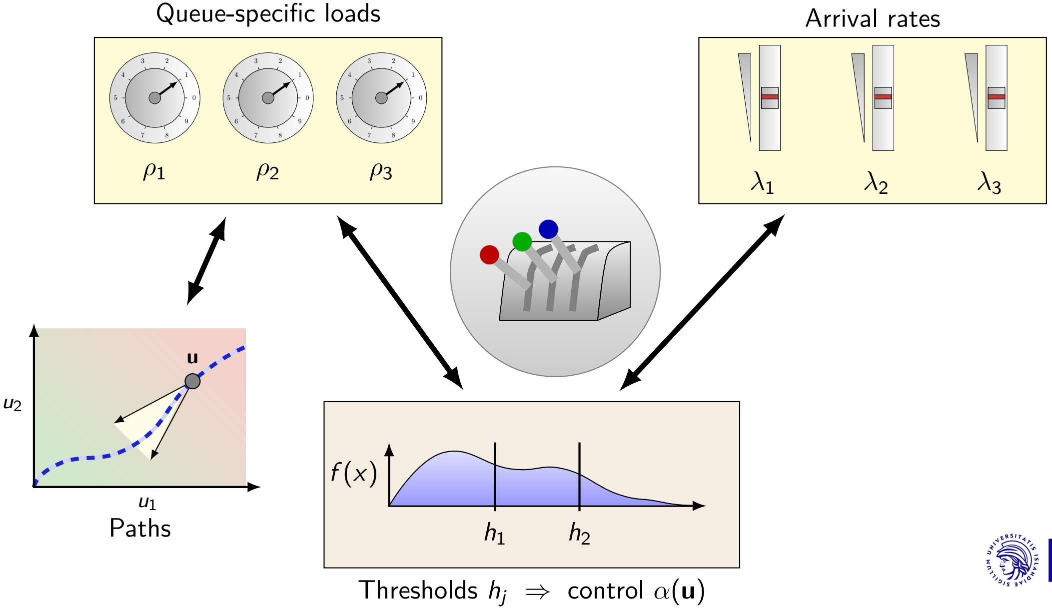
Optimal policy splits the jobsizes to n intervals using $n-1$ thresholds, $h_1 \leq h_2 \leq \ldots \leq h_{n-1}$, and routes
1) the shortest jobs to the shortest queue
2) next interval to the 2nd shortest queue, etc.



Corollary

It is sufficient to find the optimal path, which gives the ρ_j, λ_j, h_j and $\alpha(\mathbf{u})$.
That is, the policy $\alpha(\mathbf{u})$ can be defined in terms of ρ_j, λ_j or h_j.

Control Parameters



Structural Results: Scale-free Property

Theorem (Optimal paths are scale-free)

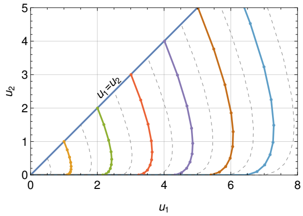
$\mathbf{p}(s)$ *optimal* $\Rightarrow \beta \cdot \mathbf{p}(s)$ *is also optimal* $\forall\, \beta > 0$.

Corollary (Monotonicity with Two Servers)

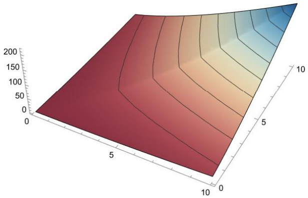
The optimal path **increases the imbalance** *as the fluid drains. Imbalance can be measured by the angle, ω, the imbalance ratio, u_2/u_1, or the relative imbalance, $(u_1 - u_2)/(u_1 + u_2)$.*

Corollary (Quadratic value function)

$$v(\beta\mathbf{u}) = \beta^2 \cdot v(\mathbf{u}) \qquad n \text{ servers}$$
$$v(\mathbf{u}) = |u|^2 \cdot w(\omega) \qquad 2 \text{ servers}$$



Optimal paths

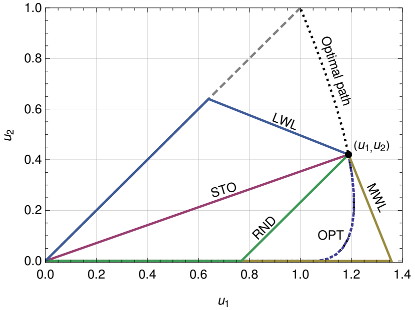


Value function

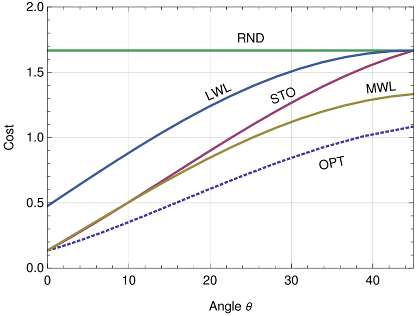
Numerical Studies

Heuristics vs. Optimal Path

- Two servers with service rates $(1/2, 1/2)$
- Exp(1)-distributed jobs and offered load $\rho = 0.7$



Paths from $\mathbf{u} = (1.2, 0.4)$

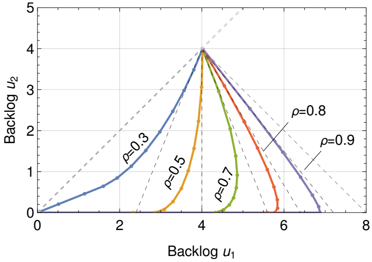


Angular value function $w(\omega)$

Imbalancing pays off!

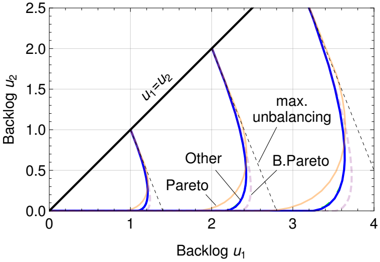
Varying Load and Jobsizes with Two Servers

- Two servers service rates $1/2$
- Exp(1)-distributed jobs
- Initial state: $\mathbf{u} = (4, 4)$



Optimal policy unbalances backlogs

- Two servers service rates $1/2$
- Offered load $\rho = 0.7$
- Different jobsize distributions

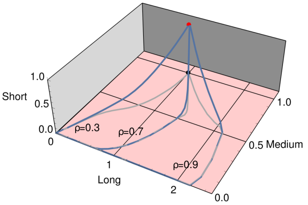


Optimal policy depends on size-distribution

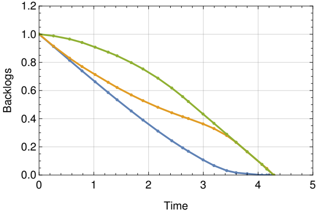
Theorem

More variable job sizes lead to lower total costs.

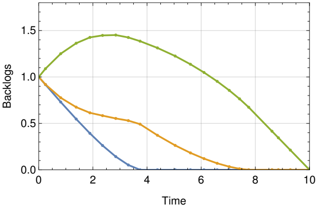
Paths with Three Servers – Varying Load



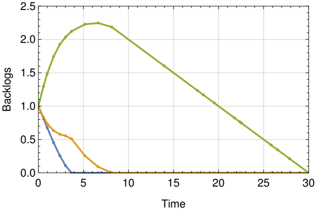
- Three servers with service rates $1/3$
- U(0,2)-distributed jobsizes
- Initial state: $\mathbf{u} = (1, 1, 1)$



$\rho = 0.3$



$\rho = 0.7$



$\rho = 0.9$

Optimal policy aggressively empties one or two servers at the expense of the third!

Conclusions

1. Fluid routing problem is an **interesting optimization problem itself!**
2. Essentially a problem in **variational calculus**
3. Many interesting **structural results**
   - More variability in job sizes *decreases* costs
   - As $\rho \to 1$, MWL is optimal
     And the mean waiting time agrees with the heavy traffic optimality results[1]
4. Gives **insight to the job dispatching problem**

Thank you! Any questions?

(esa@hi.is)

[1] R. Xie, I. Grosof, and Z. Scully, *"Heavy-traffic optimal size-and state-aware dispatching,"* Proc. of the ACM on Measurement and Analysis of Computing Systems, 2024.

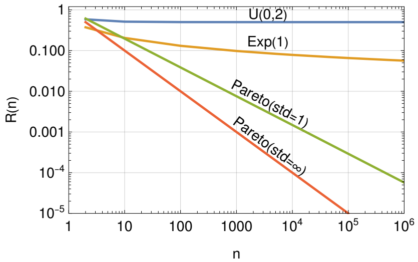
Scaling n

Comparing $\mathbb{E}[W]$ with n servers to $\mathbb{E}[W]$ with a comparable single server system gives

$$R(n) := \frac{\mathbb{E}[W_n]}{\mathbb{E}[W_1]} \to \left[1 - \Phi\left(\frac{n-1}{n}\right)\right] n \quad \text{as } \rho \to 1. \quad [\Phi(s) := F(g^{-1}(s))]$$

For large n, we have similarly

$$\boxed{R(n) \approx \frac{1}{g^{-1}(s)}.} \tag{1}$$



- With $X \sim \mathrm{U}(0, 2)$, we have $R(n) \to 1/2$; $\mathbb{E}[W]$ can decrease atmost to half!
- With Pareto distribution, $R(n) \to 0$. The rate depends on the shape parameter α.

Eq. (1) quantifies how performance scales under heavy load with different jobsize distributions!

References

[1] R. Xie, I. Grosof, and Z. Scully, *"Heavy-traffic optimal size-and state-aware dispatching,"* Proc. of the ACM on Measurement and Analysis of Computing Systems, 2024.

[2] E. Hyytiä and R. Righter, *"Towards the Optimal Dynamic Size-aware Dispatching,"* Performance Evaluation, no. 102396, 2024.

[3] E. Hyytiä, P. Jacko and R. Righter, *"Routing with too much information?,"* Queueing Systems, vol. 100, pp. 441-443, 2022.