Stochastic Optimization for Scheduling Service Systems

Brian Denton

Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI

February 20, 2024

Summary

Service System Scheduling Examples:

- Example 1: Single sever scheduling
- Example 2: Multi-server scheduling
- Example 3: Bi-criteria scheduling of multi-server, multi-stage service system

Key Take Aways

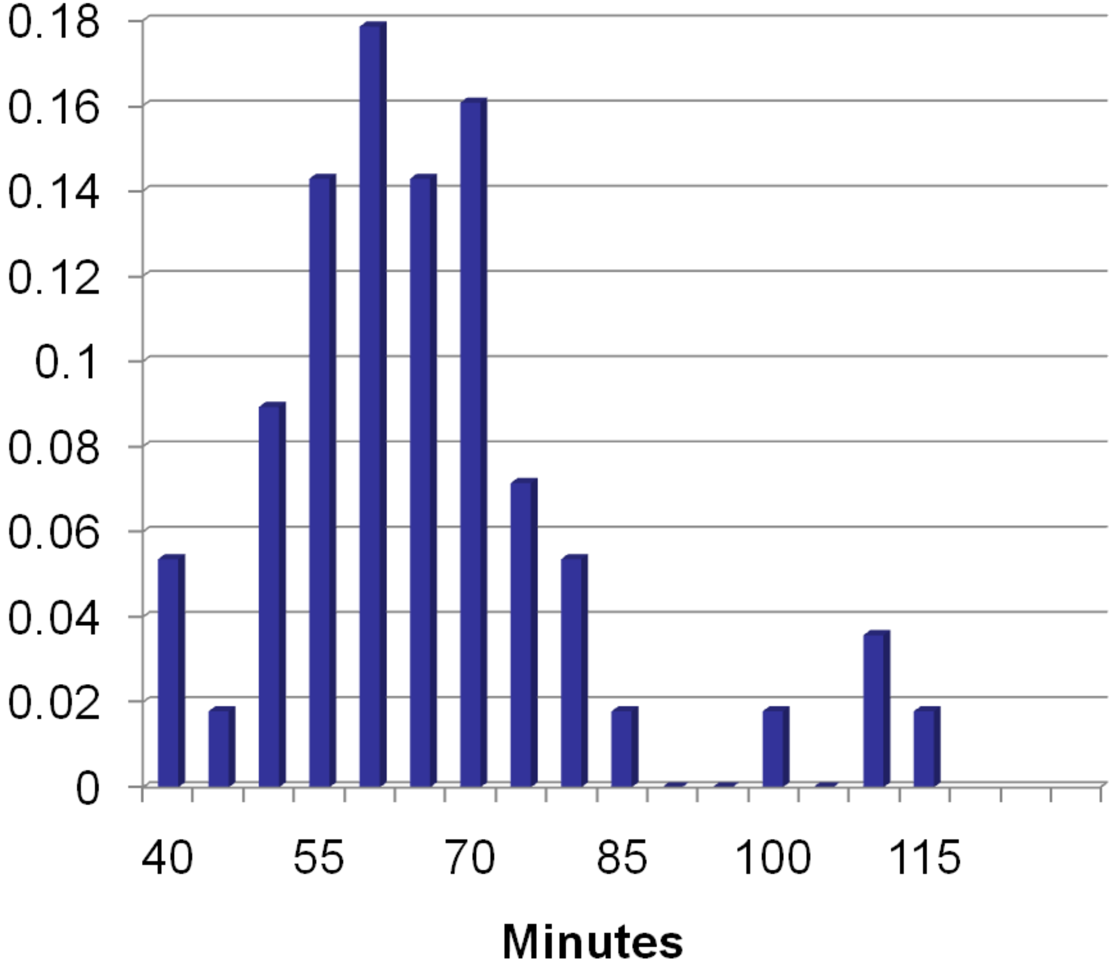
Examples of stochastic scheduling problems

- What is the optimal assignment of surgeries to operating rooms at a hospital?
- What is the optimal schedule of deliveries of raw material inventory to a manufacturer?
- What is the optimal arrival schedule of cargo ships to a port?

Complicating Factors

- High cost of customer waiting and server idling and a fixed time to complete activities
- Large number of activities to be coordinated in a constrained environment
- Uncertainty in the duration of customer service, and server availability
- Human behavior

A motivating example – surgery



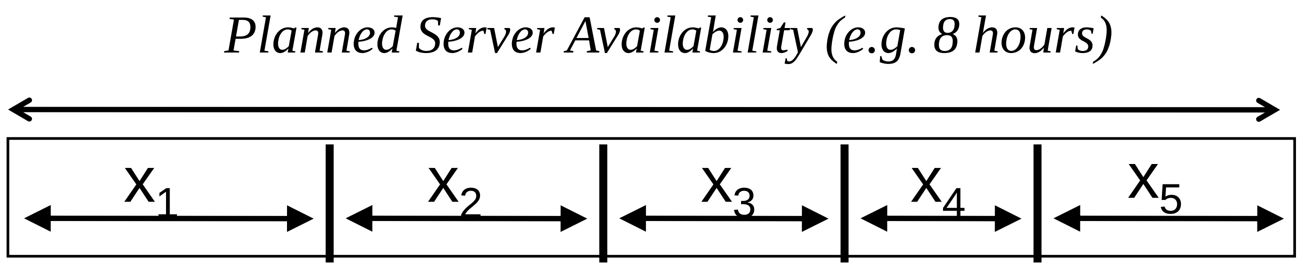
Example 1
Single Server Scheduling

Single Sever Scheduling Problem

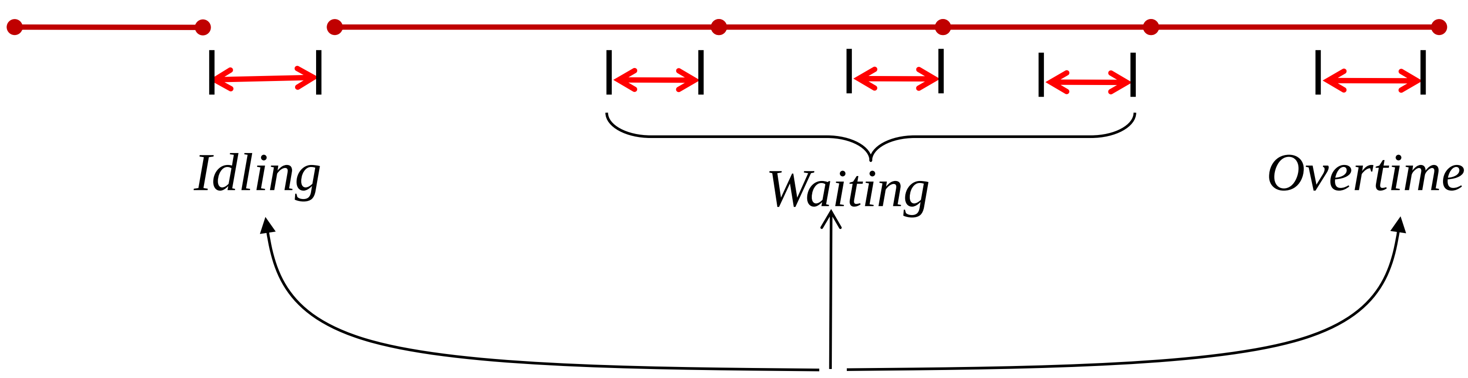
For a single server, find the optimal time to allocate for each customer to minimize the cost of:

- Customer waiting
- Server idling
- Overtime

Single Server Scheduling



Example Scenario:



Goal: Min{ Idling + Waiting + Overtime}

Stochastic Optimization Model

$$\min_{\mathrm{x}}\{\overbrace{\sum_{i=1}^{n} C_i^{w} E_{\mathbf{Z}}[W_i]}^{\text{Cost of Waiting}} + \overbrace{\sum_{i=1}^{n} C_i^{s} E_{\mathbf{Z}}[S_i]}^{\text{Cost of Idling}} + \overbrace{C^{L} E_{\mathbf{Z}}[L]}^{\text{Cost of Overtime}}\}$$

Random service time

Planned time for service

$$W_i = \max(W_{i-1} + Z_{i-1} - x_{i-1}, 0)$$

$$S_i = \max(-W_{i-1} - Z_{i-1} + x_{i-1}, 0)$$

$$L = \max(W_n + Z_n + \Sigma x_i - T, 0)$$

Literature Review – Single Server

Queuing Analysis:

- Mercer (1960, 1973)
- Jansson (1966)
- Brahimi and Worthington (1991)

Assumes steady state is reached, i.i.d. service times

Heuristics:

- White and Pike (1964)
- Soriano (1966)
- Ho and Lau (1992)

No guarantee of optimal solution

Optimization:

- Weiss (1990) – 2 customer news vendor model
- Wang (1993) – Multiple customers with phase-type distribution property
- Denton and Gupta (2003) – General stochastic programming formulation

Reformulation as a Stochastic Program

$$\min_{\mathrm{x}}\{\, E_Z[\sum_{i=2}^{n} c_i^W w_i + \sum_{i=2}^{n} c^s s_i + c^L l]\}$$

$$
\begin{aligned}
s.t.\quad & w_2 - s_2 = Z_1 - x_1 \\
& -w_2 + w_3 - s_3 = Z_2 - x_2 \\
& \qquad \ddots \qquad\qquad \ddots \\
& -w_n - s_n + l - g = Z_n - d + \sum_{j=1}^{n-1} x_i
\end{aligned}
$$

$$x_i \geq 0, w_i \geq 0, s_i \geq 0, i = 1, \ldots, n, \quad l, g \geq 0$$

Two Stage Recourse Problem

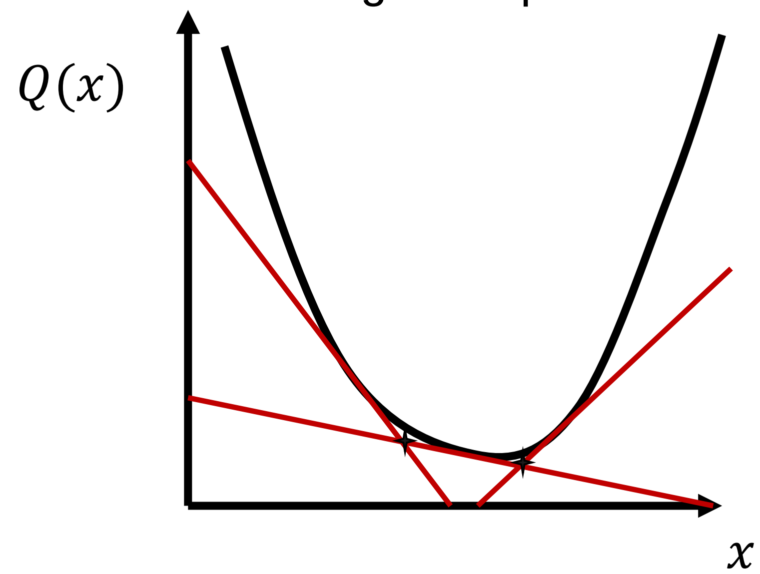
Initial Decision (x) → Uncertainty Resolved → Recourse (y)

$$\min\{\, Q(\mathbf{x}) = \sum_{k}^{K} p_k \, Q(\mathbf{x}, \mathbf{Z}^k)\}$$

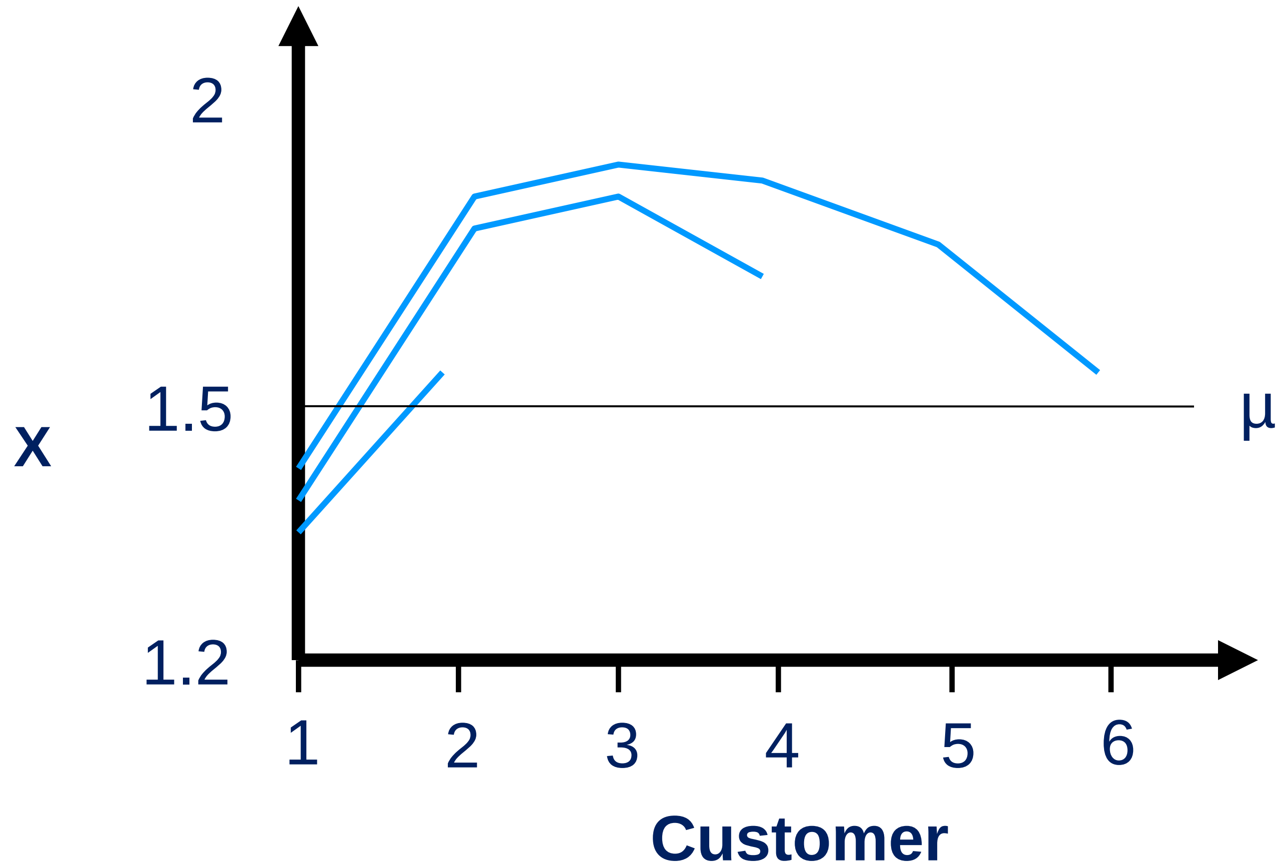
$$Q(\mathbf{x}, \mathbf{Z}^k) = \min\{\, \mathbf{c} \cdot \mathbf{y}^k \,|\, \underbrace{T\ \mathbf{x} + W\ \mathbf{y}^k}\ = \mathbf{h}^k, \mathbf{y}^k \geq 0\}$$

$$\begin{pmatrix} T & W^1 & & & \\ T & & W^2 & & \\ T & & & W^3 & \\ \vdots & & & & \ddots \\ T & & & & & W^K \end{pmatrix}$$

Solve using L-shaped method



Example: $n = 3, 5, 7$ customers with i.i.d. service times ~ $U(1,2)$, $c^w = c^s$



Insights

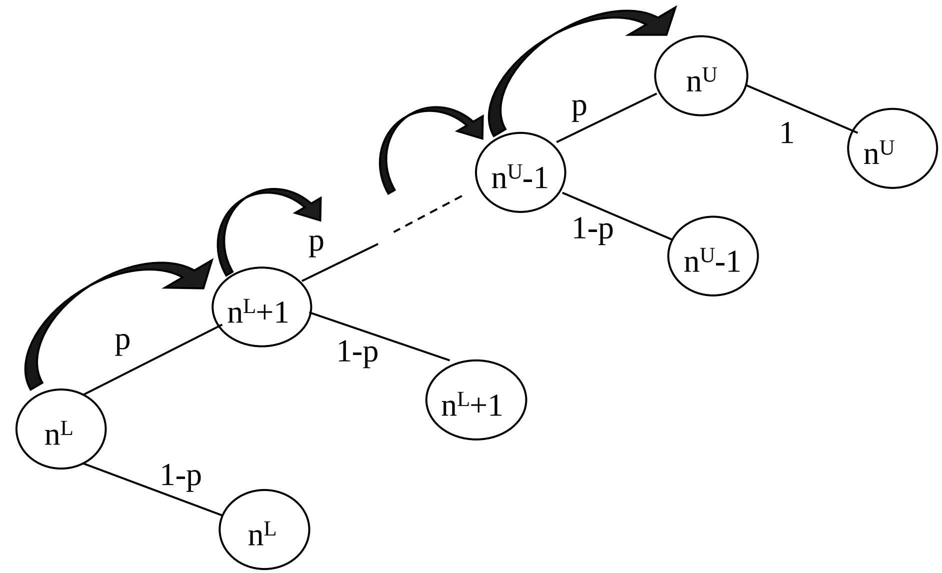
- Simple heuristics often perform poorly
- The value of the stochastic solution (VSS) can be high
- Large instances of this problem can be solved very easily

1) Denton, B.T., Gupta, D., 2003, A Sequential Bounding Approach for Optimal Appointment Scheduling, *IIE Transactions*, 35, 1003-1016

2) Denton, B.T., Viapiano, J, Vogl, A., 2007, Optimization of Surgery Sequencing and Scheduling Decisions Under Uncertainty, *Health Care Management Science,* 10(1), 13-24

There are many variations on this problem

- Customer No-shows
- Late arrivals
- Dynamic scheduling
- Endogenous uncertainty



Erdogan, S.A., Denton, B.T., "Dynamic Appointment Scheduling with Uncertain Demand," INFORMS Journal on Computing 25(1), 116-132, 2013.

Erdogan, A, Denton, B.T., Gose, "On-line Appointment Sequencing and Scheduling," IIE Transactions, 47, 1267-1286, 2015.

Zheng, Z., Denton, B.T., Xie, X., "Appointment Scheduling and the Effects of Customer Congestion on Service," *IISE Transactions*; 51(10), 1075-1090, 2019

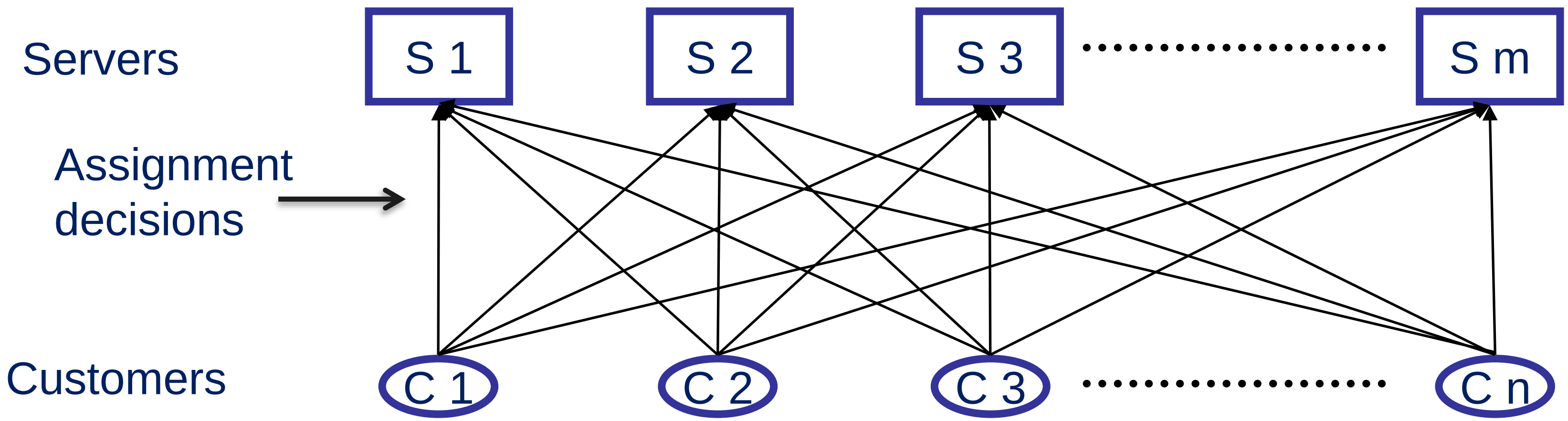
Example 2
Multiple Server Job Allocation

Multi-Server Scheduling Problem

Given a set of customers (jobs) with uncertain duration to be scheduled on a certain day decide the following:

- How many servers to make available to complete all customer service
- Which server to assign to each customer

Multi-Server Scheduling Problem



Decisions:

- How many servers to have active each day?
- Which sever to assign each job?

Extensible Bin-Packing Problem

$$x_i = \begin{cases} 1 \ \ if\ server\ i\ active \\ 0 \qquad otherwise \end{cases} \qquad y_{ij} = \begin{cases} 1 \ \ if\ customer\ j\ assigned\ to\ server\ i \\ 0 \qquad\qquad otherwise \end{cases}$$

$$Z = \min\{\sum_{i=1}^{m} c^f x_i + c^v o_i\}$$

Cost of servers + overtime

$$s.t. \qquad y_{ij} \leq x_i \quad i = 1, \ldots, m, j = 1, \ldots, n$$

Customers only scheduled to active servers

$$\sum_{i=1}^{m} y_{ij} = 1 \quad j = 1, \ldots, n$$

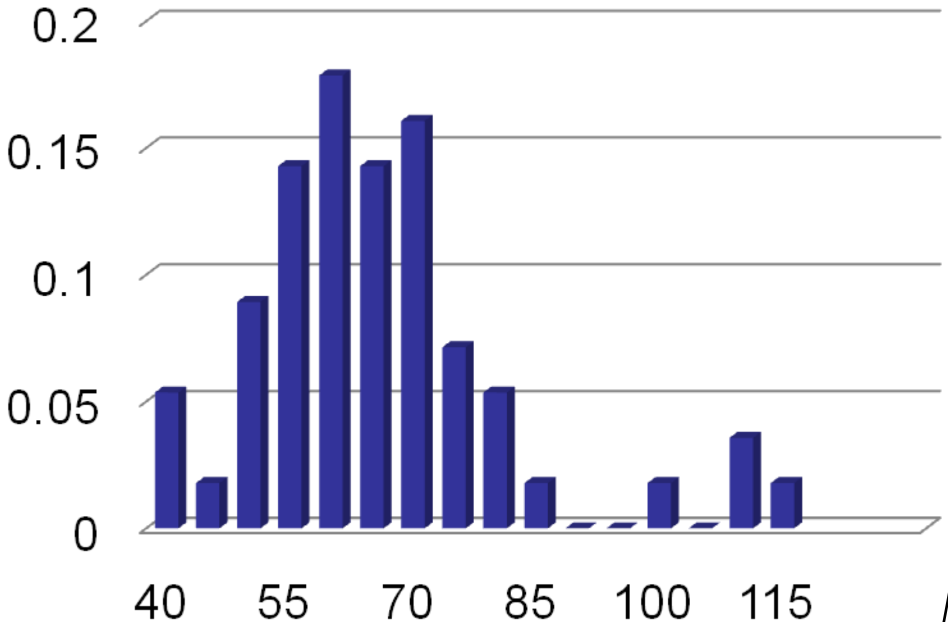
Every customer assigned to one server

$$\sum_{j=1}^{n} d_j y_{ij} - o_i \leq T x_i \quad i = 1, \ldots, m$$

Overtime if server goes past end of day of length T

$$y_{ij}, x_i\ binary, \quad o_i \geq 0$$

Two-stage stochastic mixed integer program



$$Q(\mathbf{x}) = \min\{\sum_{j=1}^{m} c^f x_j + c^v E_\omega[o_j(\omega)]\}$$

$$s.t. \quad y_{ij} \le x_j \quad \forall (i,j)$$

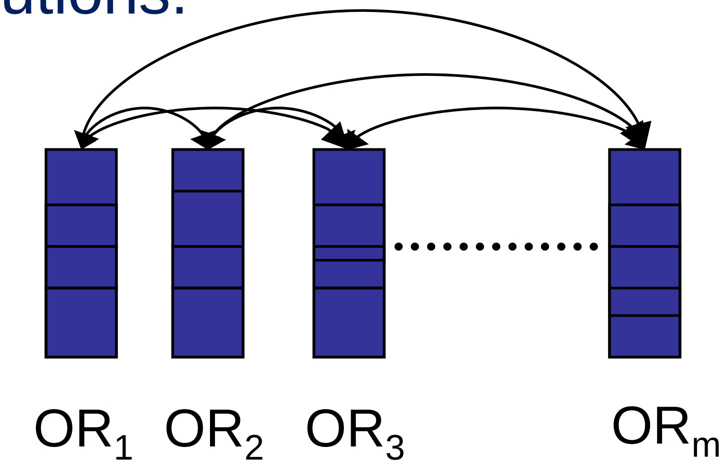
$$\sum_{j=1}^{m} y_{ij} = 1 \ \forall (i)$$

$$\sum_{i=1}^{n} d_i(\omega) y_{ij} - o_j(\omega) \le T x_j \quad \forall (i,j,\omega)$$

$$y_{ij}, x_j \in \{0,1\}, \quad o_j(\omega) \ge 0, \forall \omega$$

Symmetry is a problem

There are m! optimal solutions:



Adding the following anti-symmetry constraints reduces computation time:

$$x_1 \geq x_2$$
$$x_2 \geq x_3$$
$$\vdots$$
$$x_m \geq x_{m-1}$$

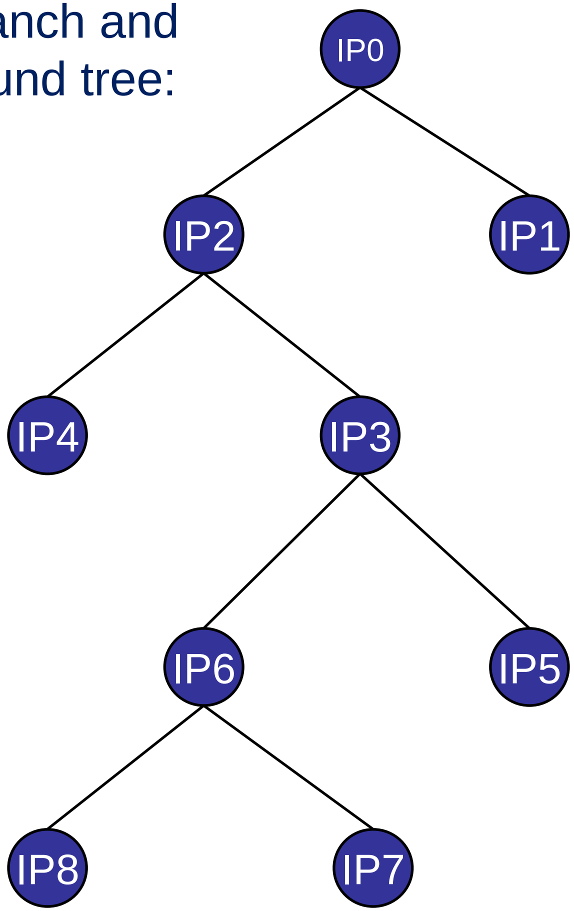
Server Ordering

$$y_{11} = 1$$
$$y_{21} + y_{22} = 1$$
$$\vdots$$
$$\sum_{j=1}^{m} y_{mj} = 1$$

Customer Assignment

Integer L-Shaped Method

Branch and bound tree:



<u>Master Problem:</u>

$$Z = \min\{\sum_{j=1}^{m} c^f x_j + \Theta\}$$

$$s.t. \quad y_{ij} \le x_j \quad \forall(i,j)$$

$$\sum_{j=1}^{m} y_{ij} = 1 \quad \forall(i)$$

$$y_{ij}, x_j \in \{0,1\}, \Theta \ge 0$$

(optimality cuts)

$$\Theta \ge E_\omega[\pi(h - Tx)]$$

Longest Processing Time First Heuristic

Sort *customers* in LPT order;

$m \leftarrow LB$ on number of *servers*;

while($o_j = 0, \forall j$)

  LPT(m);

  $m \leftarrow m + 1;$

end

Compute m^* with lowest total cost

Dell'Ollmo, Kellerer, Speranza, Tuza, *Information Processing Letters* (1998) – provides a 13/12 approximation algorithm for bin packing with a fixed number of extensible bins

Robust Formulation

Robust formulation seeks to <u>minimize</u> <u>the worst case</u> cost.

$$Z = \min\{\sum_{j=1}^{m} c^f x_j + F(x,y)\}$$

$$s.t. \quad y_{ij} \le x_j \quad \forall(i,j)$$

$$\sum_{j=1}^{m} y_{ij} = 1 \quad \forall(i)$$

$$y_{ij}, x_j \in \{0,1\} \ge 0$$

Worst case (adversary) problem

$$F(x,y) = \begin{cases} \max_{\delta}\{\sum_{j=1}^{m} \eta_j\} \\ s.t. \quad \eta_j = c_j^v \max\{0, \sum_{i:y_{ij}=1} \delta_{ij} y_{ij} - d x_j\}, \quad \forall j \\ \sum_{(i,j):y_{ij}=1}^{m} \frac{\delta_{ij} - \underline{z_i}}{\overline{z_i} - \underline{z_i}} y_{ij} \le \tau \\ \underline{z_i} \le \delta_{ij} \le \overline{z_i}, \ \forall(i,j): y_{ij} = 1 \end{cases}$$

Uncertainty budget

Results from sample test problems

Instance	1	2	3	4	5	6	7	8	9	10	Avg
LPT	.82	.97	.85	.93	.95	.85	.94	.97	.97	.92	.92
MV	.81	.95	.85	.92	.90	.86	.93	.89	.96	.86	.90
Robust	.93	.97	.97	.92	.89	.94	.92	.90	.97	.92	.92

Table 1: Cost of 0.5 hours overtime equals cost, c^f, of adding a server

Instance	1	2	3	4	5	6	7	8	9	10	Avg
LPT	1.0	1.0	1.0	1.0	1.0	.99	.99	.97	.99	1.0	.99
MV	1.0	1.0	1.0	1.0	.99	.99	.97	.97	.98	1.0	.99
Robust	.95	1.0	.95	.93	.94	.88	.97	.99	.96	.90	.95

Table 2: Cost of 2 hours overtime equals cost, c^f, of adding a server

LPT = longest processing time first heuristic, MV = mean value problem, Robust = solution to robust integer program. Results expressed as the ratio of optimal solution to solution generated by MV, LPT, Robust

Insights

- LPT works well when overtime costs are low and it has a favorable *performance ratio*

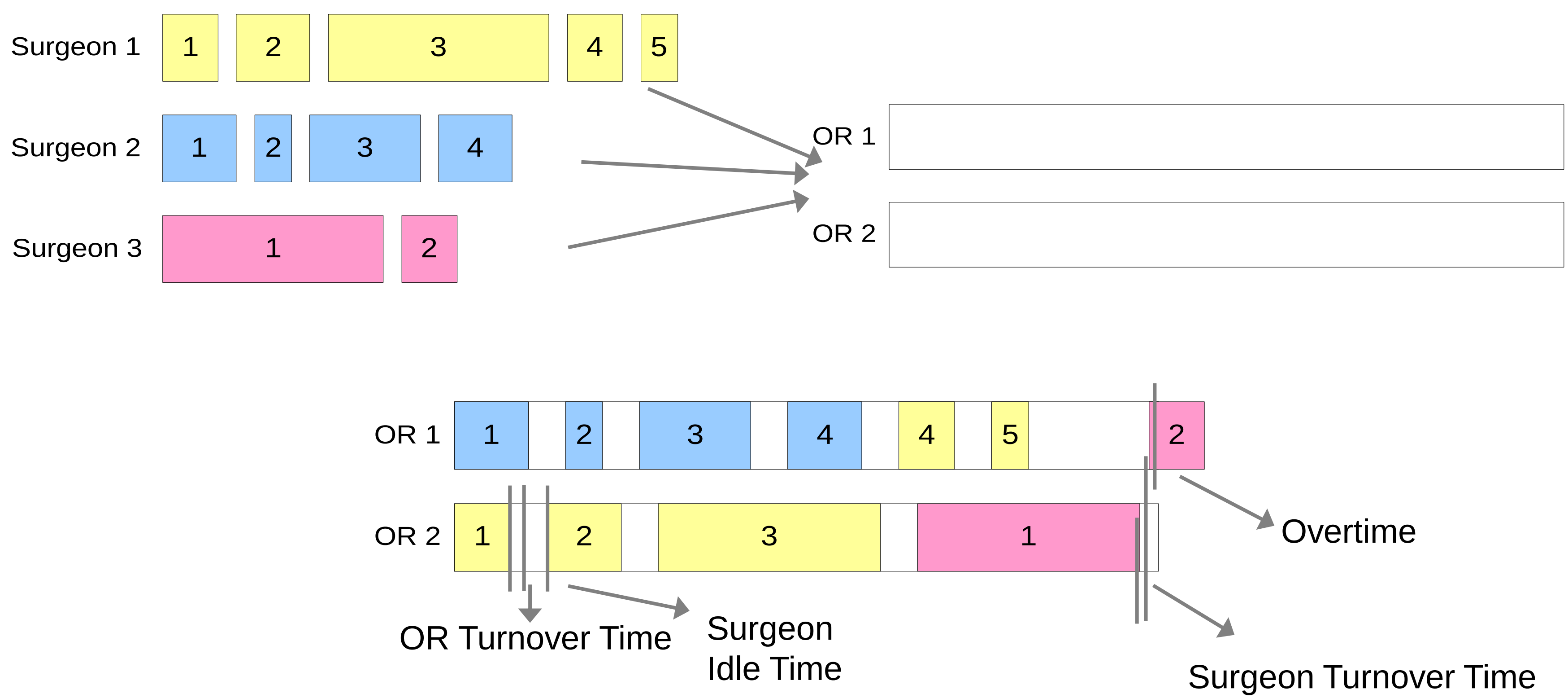
- LPT is better (and much easier) than solving MV problem in most cases

- Robust IP is better than LPT when overtime costs are high

Denton, B.T., Miller, A., Balasubramanian, H., Huschka, T., 2010, Optimal Surgery Block Allocation Under Uncertainty, *Operations Research* 58(4), 802-816, 2010

Zheng, Z., Denton, B.T., Xie, X., “Branch-and-Price for Chance-Constrained Bin Packing,” *INFORMS Journal on Computing*; 32(3):547-564, 2020

Relaxing assumptions about assignment decisions leads to challenging problems



Batun, S., Denton, B.T., Huschka, T.R., Schaefer, A.J., The Benefit of Pooling Operating Rooms Under Uncertainty, *INFORMS Journal on Computing*, 23(2), 220-237, 2012.

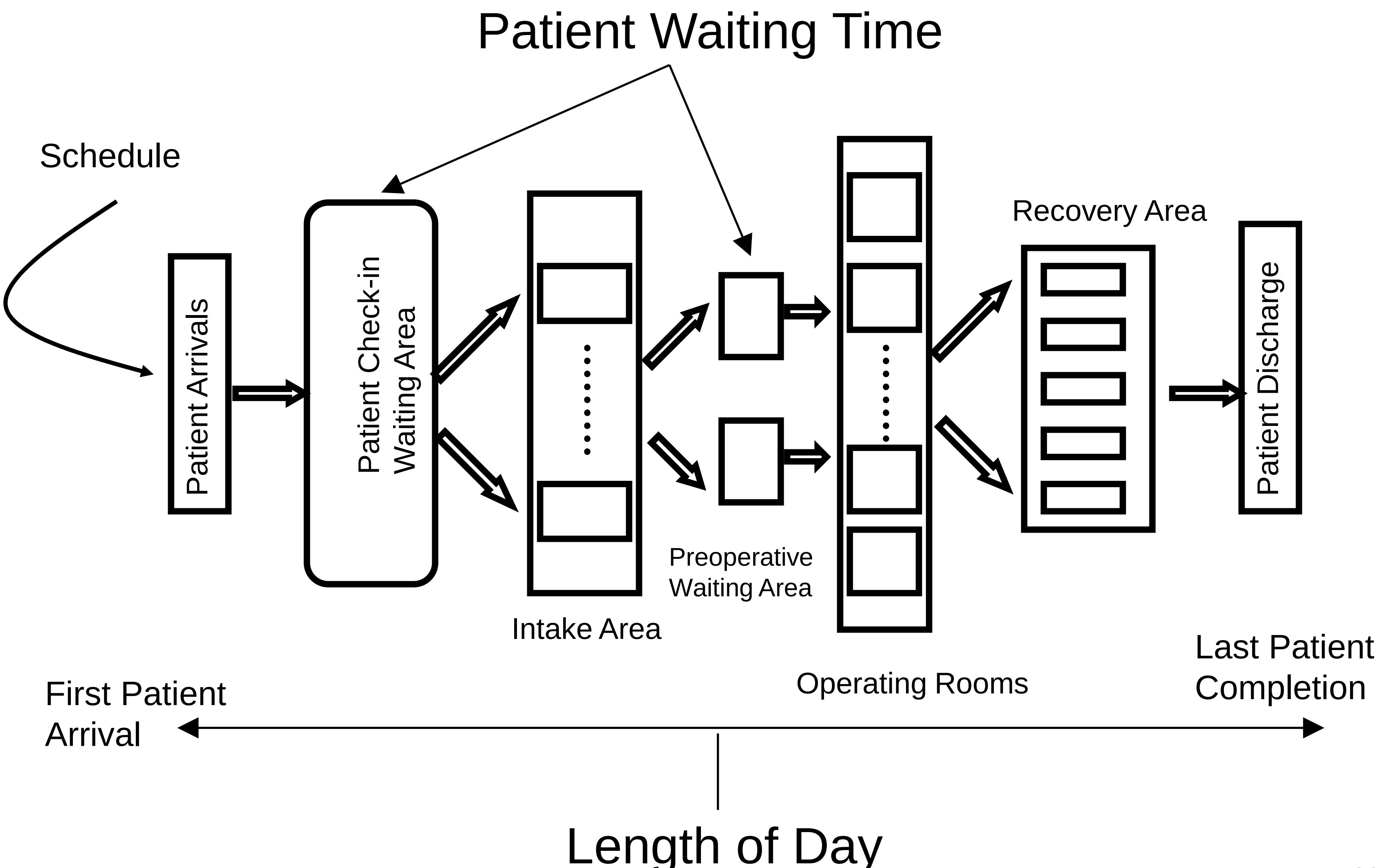
Example 3
Multi-Stage Service System

Multi-Stage Server Scheduling Problem

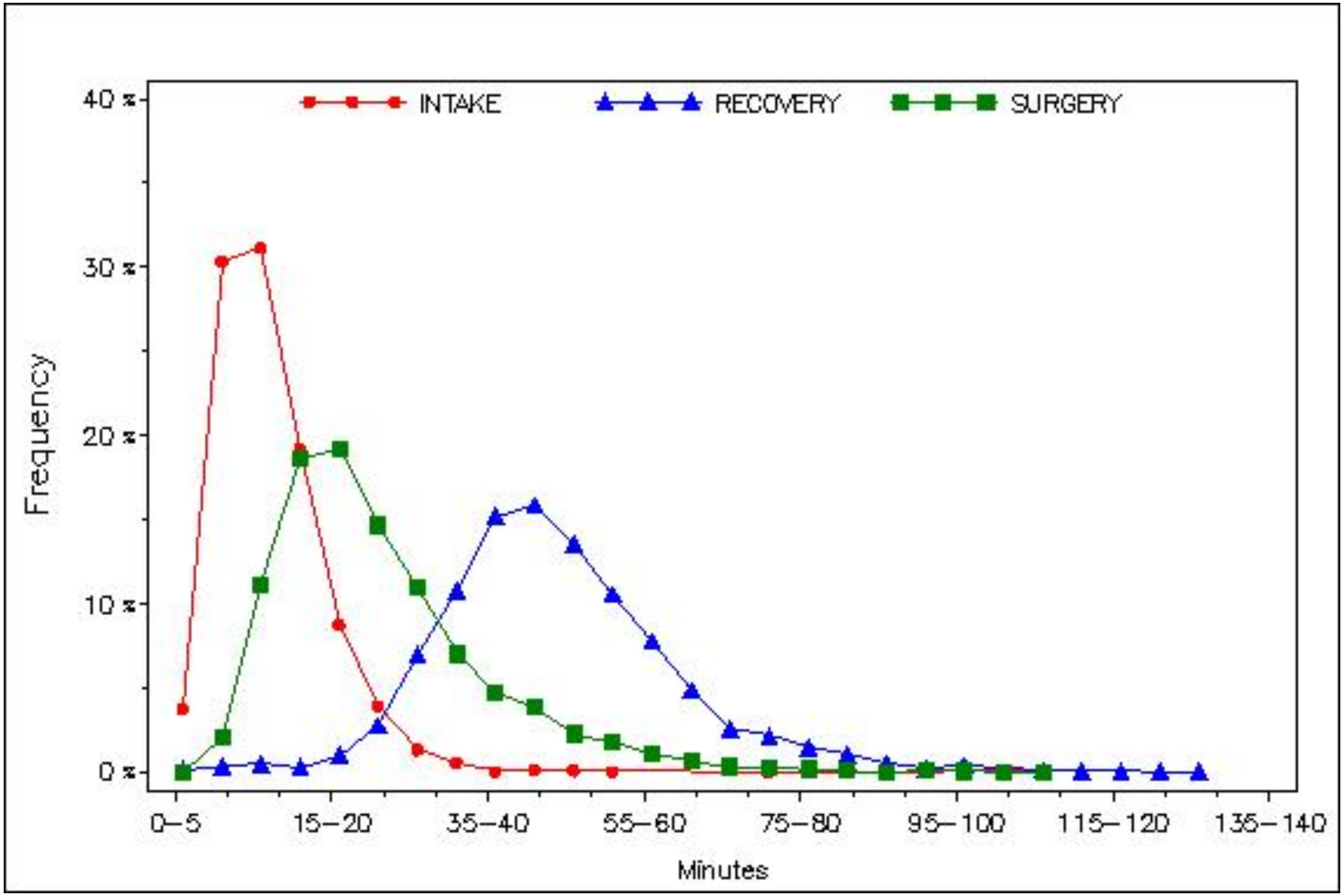
Find the Pareto optimal appointment times for patients having a procedure in an ambulatory surgery center to trade-off:

- Expected patient waiting
- Expected length of day

Context: Outpatient Procedure Centers



Intake, Procedure and Recovery Distributions



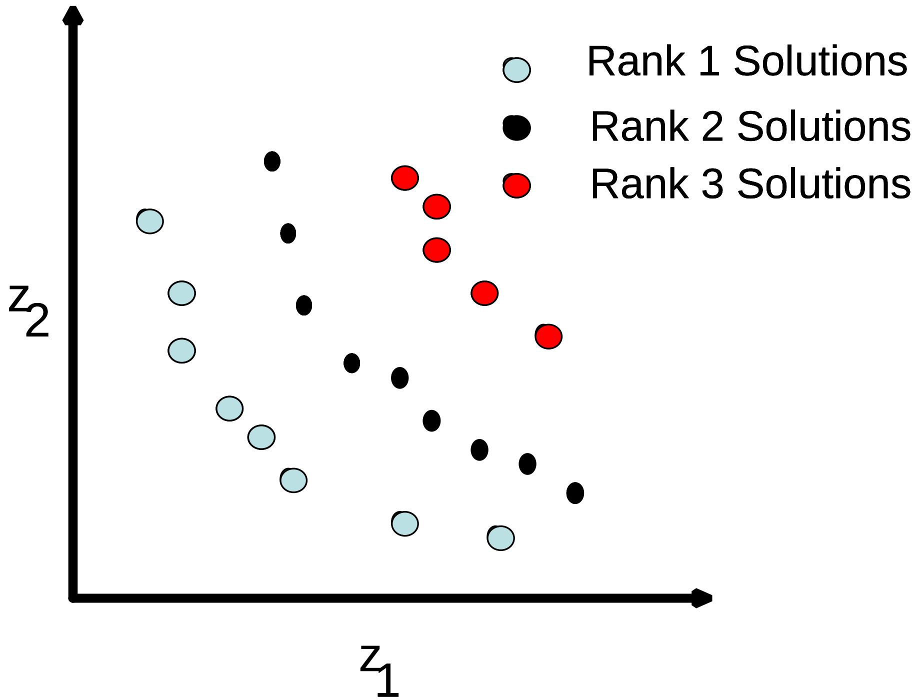
Simulation-optimization

Decision variables: scheduled arrival times to be assigned to *n* patients each day

Goal: Generate Pareto optimal schedules to understand tradeoffs between patient waiting and length of day

- Schedules generated using a genetic algorithm (GA)
- Non-dominated sorting used to identify the Pareto set and feedback into GA

The non-dominated sorting genetic algorithm (NSGA-II) of Deb *et al.*(2000):



Selection Procedure

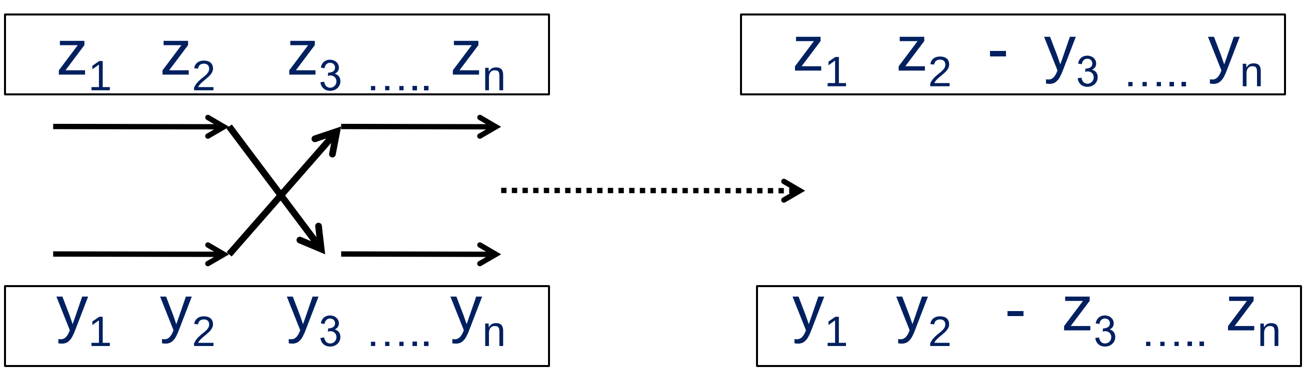
Sequential two stage indifference zone ranking and selection procedure of Rinott (1978) to compute the number of samples necessary to determine whether a solution *i* “dominates” *j*

Solution *i* “dominates” *j* if:

$$E[W_i] < E[W_j] \quad \text{and} \quad E[L_i] < E[L_j]$$

Genetic Algorithm

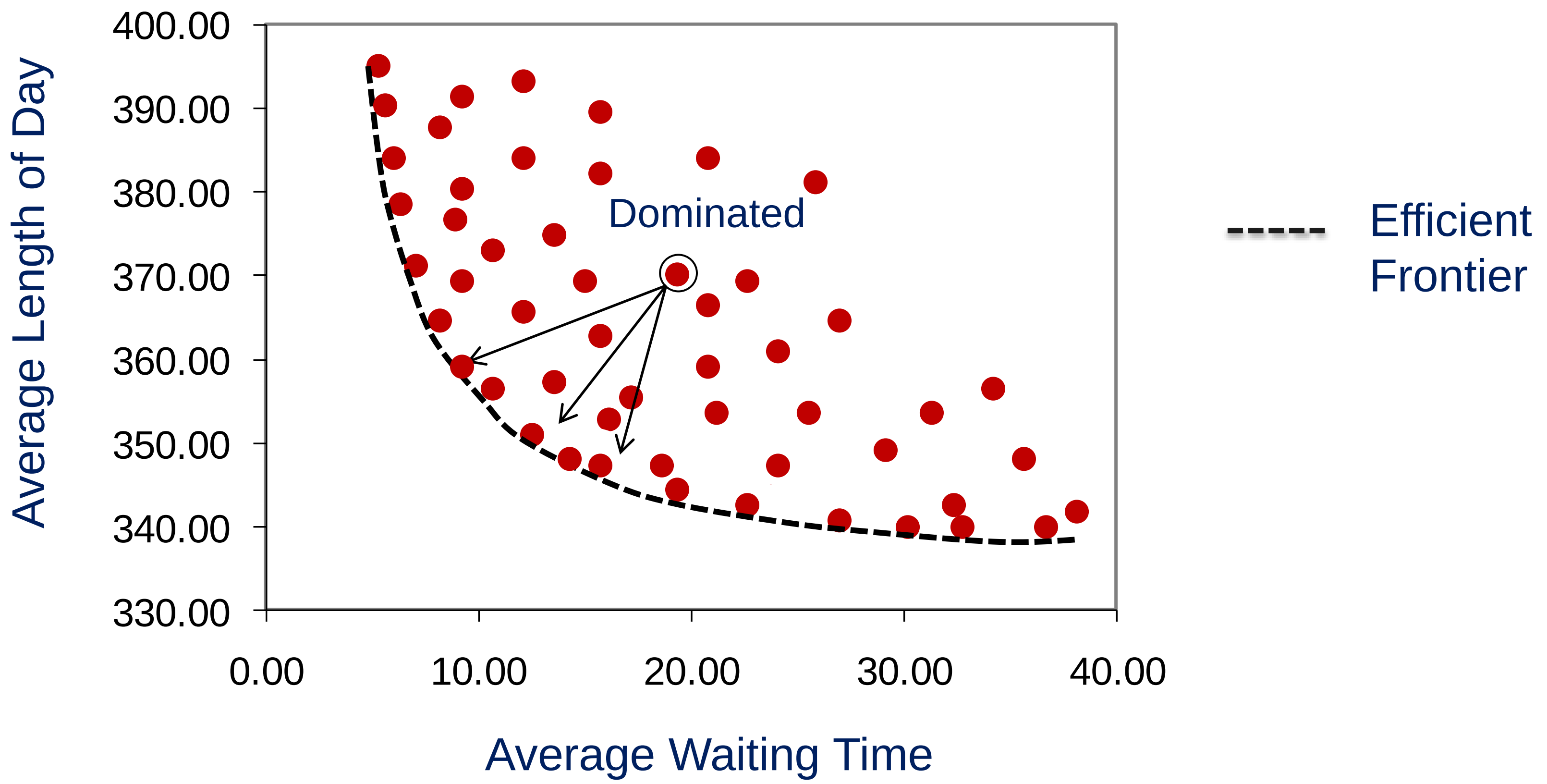
- Randomly generated initial population of schedules
- Selection based on 1) ranks and 2) crowding distance
- Mutation
- Single point crossover:



Parents

Children

Schedule Optimization



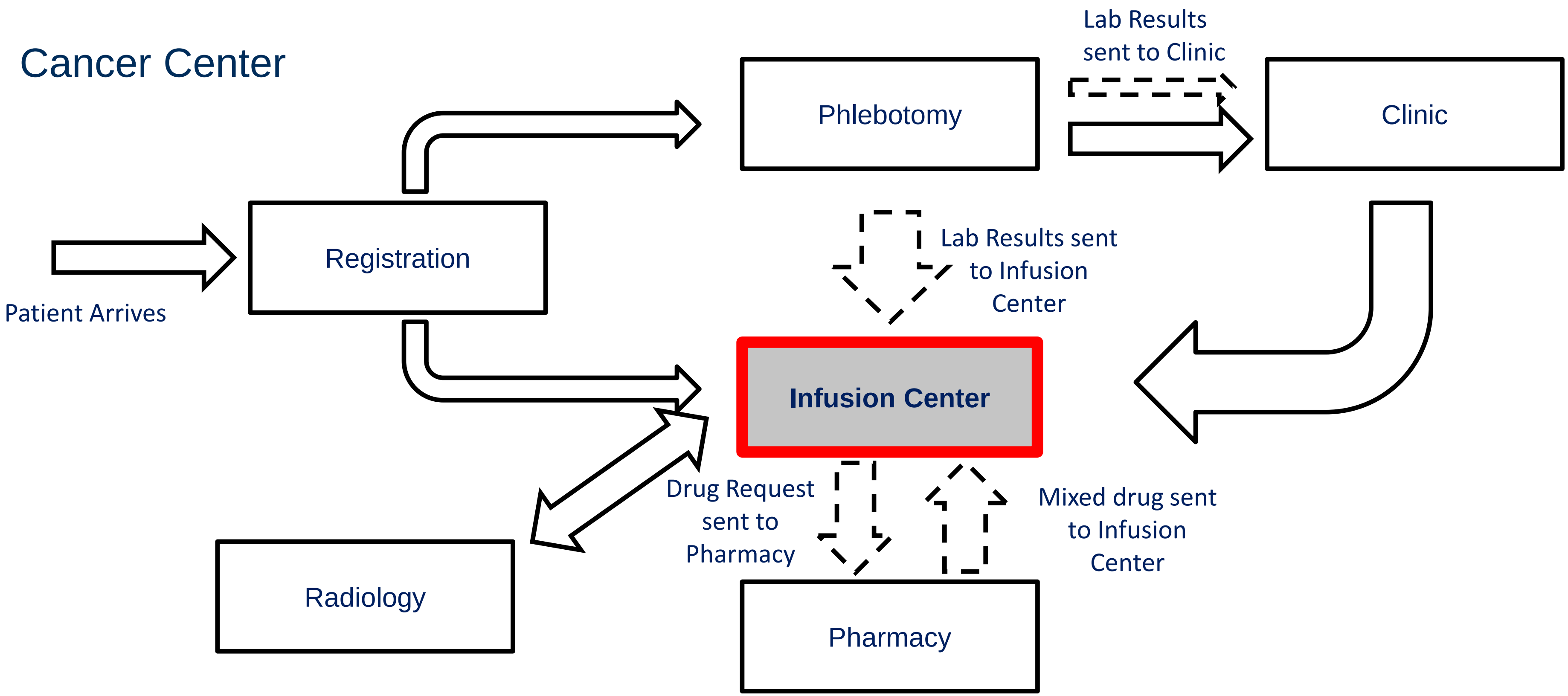
Insights

- A simple simulation-optimization approach provides significant improvement to schedules used in practice

- Substantial reduction in average waiting time is possible with a very limited increase in average length of day

Gul, S., Denton, B.T., Fowler, J., 2011 Bi-Criteria Scheduling of Surgical Services for an Outpatient Procedure Center, *Production and Operations Management*, 20(3), 406-417

Many service systems have complex interactions



Woodall, Jonathan C., Tracy Gosselin, Amy Boswell, Michael Murr, and Brian T. Denton. "Improving patient access to chemotherapy treatment at Duke Cancer Institute." *Interfaces* 43, no. 5 (2013): 449-461.

Key Takeaways

- Modeling uncertainty often matters!

- Stochastic scheduling problems can be hard, but a special structure often exists to exploit.

- Stochastic optimization is a powerful tool for scheduling in many contexts

Acknowledgements

Hari Balasubramanian (University of Massachusetts)
Sakine Batun (METU, Turkey)
Maya Bam (General Motors)
Bjorn Berg (University of Minnesota)
Ayca Erdogan (San Jose State University)
Serhat Gul (TED University, Turkey)
Todd Huschka (Mayo Clinic)
Andrew Miller (FedEx)
Heidi Nelson (Mayo Clinic)
Andrew Schaefer (Rice University)
Zheng Zhang (Zhejiang University)

Some of this work was funded in part by grants from the *Service Enterprise Systems* program at the *National Science Foundation*

Thank You!

Brian Denton
University of Michigan

btdenton@umich.edu

These slides and the papers cited can be found at my website:

http://umich.edu/~btdenton