Simulation Cutting Approach: Joint Workstation, Workload and Buffer Allocation Problem

Mengyi Zhang Shanghai Jiao Tong University

Andrea Matta Politechnico di Milano

Arianna Alfieri Politechnico di Torino

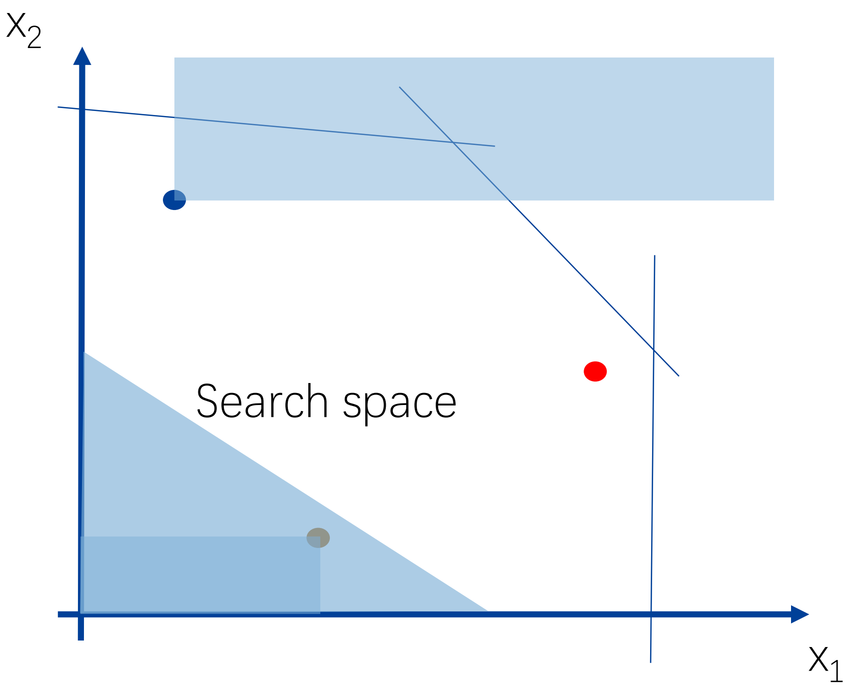
Giulia Pedrielli Arizona State University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Research framework

- The main goal of the research is to develop a methodology (**Simulation cutting approach**) to reduce the search space in simulation-optimization problems.

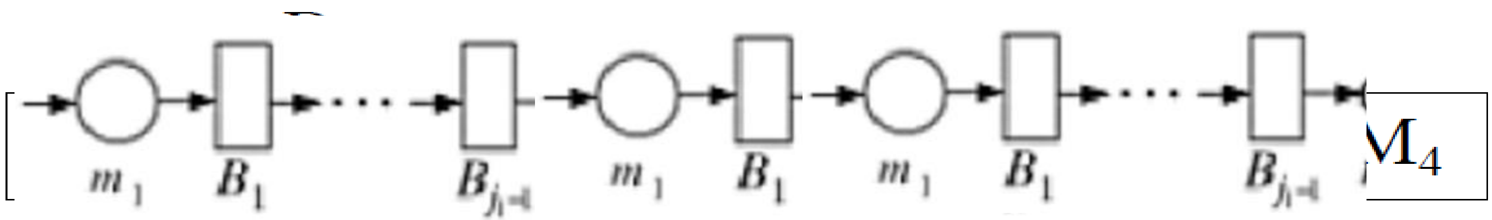


The features we want:

- A structure behind simulation output
- General enough to be used in couple with optimization techniques
- Flexible enough to be customized on the problem to exploit structural properties

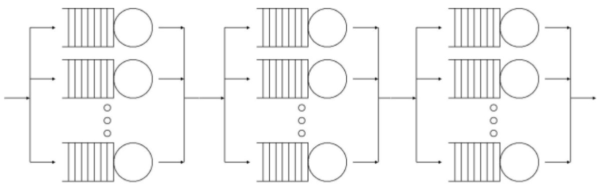
Problem Statement

- The Joint Workstation, Workload and Buffer Allocation Problem (JWWBAP) is a production system design problem:



  - **Assumptions**: stochastic processing time, general distributions, continuously divided workload, finite buffer capacity, known expected total processing time.
  - **Decision variables**: number of workstations m, workload s_j, buffer capacity b_j.
    - Workload $s_j = \frac{Expected\ processing\ time\ at\ workstation\ j}{Expected\ total\ processing\ time}$
  - **Objective**: minimize the investment cost.
  - **Constraint**: a target throughput α^* must be satisfied.

Problem related literature



Three kinds decision variables (Hillier F S et al.1995)**:**

- **Number of servers at each station**
- **Service rate of the servers**
- **Buffer capacity**

Literature	Server number	Service rate	Buffer capacity	Optimization approach	Evaluation approach
Shanthikumar J.G et al. 1987	X		X	Analytical method (concave function)	
Hillier F S et al. 1995	X	X	X	Enumeration Parallel tangents	Analytical method
Spinellis D et al. 2000	X	X	X	simulated annealing algorithm	Analytical method
Horng S C et al. 2016		X	X	elitist teaching-learning-based optimization and optimal computing budget allocation methods	Meta model
Van Woensel T et al. 2010	X		X	Non-linear optimization methodology	Analytical method
Smith, J.G 2016		X	X	Mixed-integer sequential quadratic programming	Analytical method

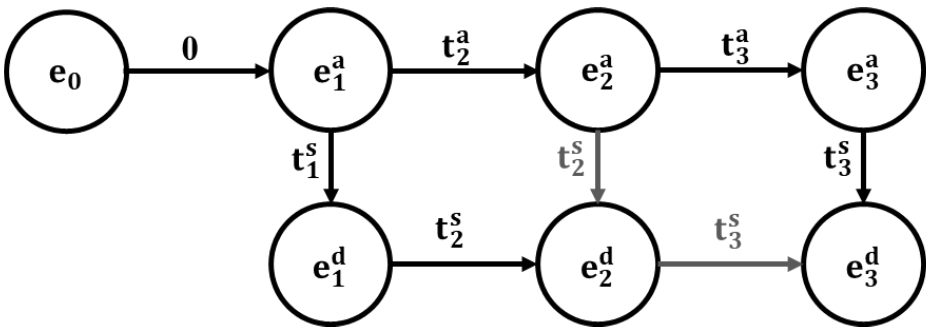
Discrete Events Optimization

- **DEO** is an integrated simulation optimization modeling framework.
- A DEO model can describe the simulation trajectories of a set of possible systems, i.e., its configuration is defined with variables.
- A DEO model is a Mathematical Programming (MP) model of Event Relationship Graphs (ERGs).
- DEO main features:
  - Simulation is regarded as a **white box**.
  - Ability to solve stochastic system optimization problems.
  - Introducing MP solution techniques (e.g., **Benders Decomposition (BD)**).

Example of DEO: G/G/1

G/G/1 with infinite buffer and 3 entities

Decision variables:
average inter arrival time, average service time



ERG: Entity Relationship Graph (Schruben, 1983)

$$min\{c_s t^s - p_a t^a + N_\epsilon \epsilon + \sum_{i=1}^{3}(e_i^a + e_i^d)\}$$

s.t.

$$t_i^a = t^a z_i^a$$

$$t_i^s = t^s z_i^s$$

$$e_1^a = 0$$

$$e_2^a - e_1^a \geq t_2^a$$

$$e_3^a - e_2^a \geq t_3^a$$

$$e_1^d - e_1^a \geq t_1^s$$

$$e_2^d - e_2^a \geq t_2^s$$

$$e_3^d - e_3^a \geq t_3^s$$

$$e_2^d - e_1^d \geq t_2^s$$

$$e_3^d - e_2^d \geq t_3^s$$

$$\frac{\sum_{i=1}^{3}(e_i^d - e_i^a - t_i^s)}{3} - \epsilon \leq \tau^*$$

DEO related literature

DEO: Generalized modeling methodology
Pedrielli (2013), Matta et al. (2014), Pedrielli et al. (2015a), Pedrielli at al. (2017)

	Approximate solution	Exact solution
BAP	*Matta (2008)* *Alfieri et al. (2012)* *Matta et al.(2015b)* Göttlich et al. (2016)	*Stolletz and Weiss (2013)* *Stolletz and Weiss (2015)* *Weiss et al. (2017)*
JWWBAP	*Zhang et al. (2016)*	*This work*
Production rate control problem		*Tan (2015)*

$$min\left\{\underbrace{C_M\sum_{j=1}^{U_M} m_j}_{\text{Workstation cost}} + \underbrace{C_B\sum_{j=1}^{U_M-1}\sum_{k=1}^{U_B} kx_{jk}}_{\text{Buffer cost}} + \underbrace{\sum_{j=1}^{U_M}\sum_{i=1}^{N} e_{ij}^f}_{\text{Simulation}} + \underbrace{N_\epsilon \epsilon}_{\text{Feasibility}}\right\}$$

s.t.

$$\sum_{j=1}^{U_M} s_j = 1$$
$$s_j \le m_j, \forall j$$
$$m_j \le m_{j-1}, \forall j$$
$$\sum_{k=1}^{U_B} x_{jk} = 1, \forall j$$
$$t_{ij} = \phi(s_j, z_{ij}), \forall i, j$$
$$e_{i1}^f \ge e_i^a + t_{i1}, \forall i, j$$
$$e_{ij}^f - e_{i-1,j}^f \ge t_{i,j}, \forall i, j$$
$$e_{ij}^f - e_{i,j-1}^f \ge t_{i,j}, \forall i, j$$
$$e_{ij}^f - e_{i-k,j+1}^f \ge t_{ij} - M(1 - x_{jk}), \forall i, j, k$$
$$\frac{\sum_{j=1}^{U_M}\sum_{i=D}^{N} e_{ij}^f}{N - D} - \epsilon \le \alpha^*$$
$$m_j, x_{jk} \in \{0,1\}, 0 \le s_j \le 1$$
$$e_{ij}^f \ge 0, t_{ij} \ge 0, \epsilon \ge 0$$

Parameters	
U_M	Upper bound of workstation number
C_M	Unit cost of one workstation
U_B	Upper bound of stage buffer capacity
C_B	Unit cost of one buffer slot
N	Number of parts in simulation
N_ϵ	Penalty for violence of target throughput
z_{ij}	Random numbers for stochastic processing time generation
M	Large number in big-M constraints
α^*	Target average inter-departure time
D	Number of parts in the warm-up period

Variables	
m_j	Workstation allocation Workstation number=$\sum_{j=1}^{U_M} m_j$
s_j	Workload allocation
x_{jk}	Buffer allocation Buffer capacity $b_j = \sum_{k=1}^{U_B} kx_{jk}$
e_{ij}^f	Finishing time of part i at stage j
t_{ij}	Processing time of part i at stage j
ϵ	Feasibility gap variable

$$min\left\{C_M\sum_{j=1}^{U_M}m_j+C_B\sum_{j=1}^{U_M-1}\sum_{k=1}^{U_B}kx_{jk}+\sum_{j=1}^{U_M}\sum_{i=1}^{N}e_{ij}^f+N_\epsilon\epsilon\right\}$$

s.t.

Optimization	$\sum_{j=1}^{U_M}s_j=1$	Workload is completely allocated
	$s_j\le m_j,\forall j$	Workload s_j>0 ⇔ workstation j is allocated
	$m_j\le m_{j-1},\forall j$	Workstations are arranged in a flow
	$\sum_{k=1}^{U_B}x_{jk}=1,\forall j$	Only one size is allocated to each buffer
	$t_{ij}=\phi(s_j,z_{ij}),\forall i,j$	Random variate generation
Simulation	$e_{i1}^f\ge e_i^a+t_{i1},\forall i,j$	Parts arrive before processing
	$e_{ij}^f-e_{i-1,j}^f\ge t_{i,j},\forall i,j$	Part sequence
	$e_{ij}^f-e_{i,j-1}^f\ge t_{i,j},\forall i,j$	Processing sequence
	$e_{ij}^f-e_{i-k,j+1}^f\ge t_{ij}-M(1-x_{jk}),\forall i,j,k$	Blocking due to finite buffer
	$\frac{\sum_{j=1}^{U_M}\sum_{i=D}^{N}e_{ij}^f}{N-D}-\epsilon\le\alpha^*$	Performance constraint
	$m_j,x_{jk}\in\{0,1\},0\le s_j\le 1$	
	$e_{ij}^f\ge 0,t_{ij}\ge 0,\epsilon\ge 0$	

The complexity of the exact model is high.

Processing time generation

$$t_{ij} = \phi(s_j, z_{ij})$$

z_{ij}: random generated number for processing time of part *i* at station j (known)
s_j: workload at station *j*
ϕ: a linear function of s_j

Some examples

$$t_{ij} = z_{ij} s_j$$

- Beta distribution:
  - $t_{ij} \sim Beta(2,2)$ on $(0, 2Ts_j)$, $z_{ij} \sim Beta(2,2)$ on $(0, 2T)$
- Exponential distribution:
  - $t_{ij} \sim Exp(\frac{1}{Ts_j})$, $z_{ij} = -T ln(u_{ij})$, $u_{ij} \sim unif(0,1)$

$$min\left\{C_M\sum_{j=1}^{U_M} m_j + C_B\sum_{j=1}^{U_M-1}\sum_{k=1}^{U_B} kx_{jk} + \sum_{j=1}^{U_M}\sum_{i=1}^{N} e_{ij}^f + N_\epsilon \epsilon\right\}$$

s.t.

Optimization:
the master problem

$$\sum_{j=1}^{U_M} s_j = 1$$

$$s_j \le m_j, \forall j$$

$$m_j \le m_{j-1}, \forall j$$

$$\sum_{k=1}^{U_B} x_{jk} = 1, \forall j$$

$$t_{ij} = \phi\left(s_j, z_{ij}\right), \forall i, j$$

Simulation:
the subproblem

$$e_{i1}^f \ge e_i^a + t_{i1}, \forall i, j$$

$$e_{ij}^f - e_{i-1,j}^f \ge t_{i,j}, \forall i, j$$

$$e_{ij}^f - e_{i,j-1}^f \ge t_{i,j}, \forall i, j$$

$$e_{ij}^f - e_{i-k,j+1}^f \ge t_{ij} - M\left(1 - x_{jk}\right), \forall i, j, k$$

$$\frac{\sum_{j=1}^{U_M}\sum_{i=D}^{N} e_{ij}^f}{N-D} - \epsilon \le \alpha^*$$

$$m_j, x_{jk} \in \{0,1\}, 0 \le s_j \le 1$$

$$e_{ij}^f \ge 0, t_{ij} \ge 0, \epsilon \ge 0$$

Benders Decomposition

Original problem

$$min\{\mathbf{c}^{\mathrm{T}}\boldsymbol{x} + \mathrm{f}(\boldsymbol{y}) + \boldsymbol{N}^{\boldsymbol{T}}\boldsymbol{\epsilon}\}$$
$$s.t.\ \ \mathbf{A}\boldsymbol{x} + \mathbf{F}(\boldsymbol{y}) + \boldsymbol{\epsilon} \geq \mathbf{b}$$

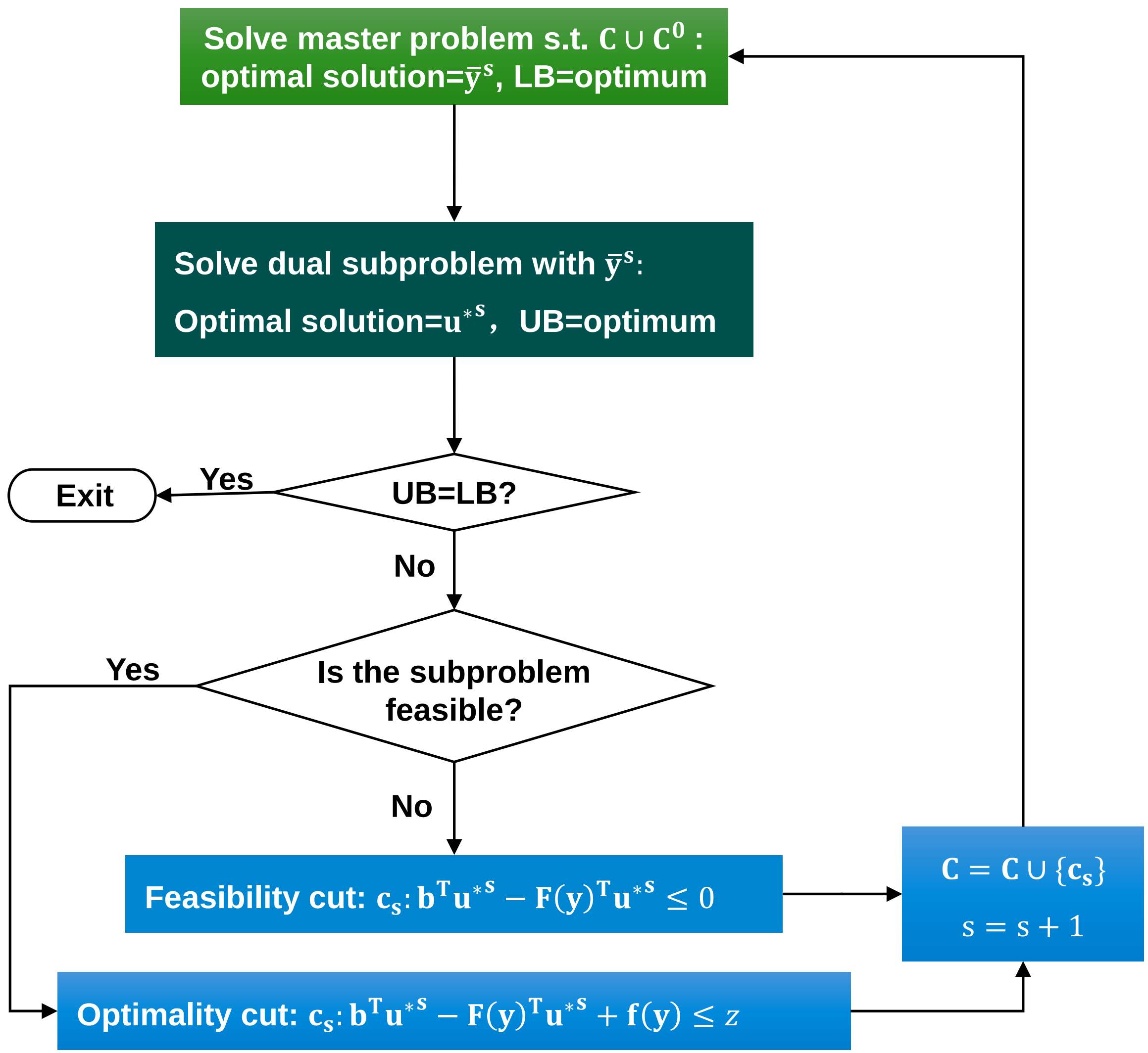
Subproblem

$$min\{\mathbf{c}^{\mathrm{T}}\boldsymbol{x} + \mathbf{N}^{\mathbf{T}}\boldsymbol{\epsilon} + \mathrm{f}(\bar{\boldsymbol{y}})\}$$
$$s.t.\ \ \mathbf{A}\boldsymbol{x} + \mathbf{F}(\bar{\boldsymbol{y}}) + \boldsymbol{\epsilon} \geq \mathbf{b}$$

Dual subproblem

$$max\{\mathbf{u}^{\mathrm{T}}(\mathbf{b} - \mathbf{F}(\bar{\boldsymbol{y}})) + \mathrm{f}(\bar{\boldsymbol{y}})\}$$
$$s.t.\ \ \mathbf{u}^{\mathbf{T}}\mathbf{A} \leq \mathbf{c}$$
$$\boldsymbol{\theta} \leq \boldsymbol{N}$$

Number of iterations: s$= 0$
Set of generated cuts: $\mathrm{C} = \emptyset$
Set of initial constraints: $\mathrm{C}^{\mathbf{0}}$



The subproblem

$$min\left\{\sum_{j=1}^{U_M}\sum_{i=1}^{N} e_{ij}^{f} + N_\epsilon \epsilon\right\}$$

$$
\begin{array}{ll}
e_{i1}^{f} \geq e_i^{a} + t_{i1} & : a_i \\
e_{ij}^{f} - e_{i-1,j}^{f} \geq t_{i,j} & : u_{ij} \\
e_{ij}^{f} - e_{i,j-1}^{f} \geq t_{i,j} & : v_{ij} \\
e_{ij}^{f} - e_{i-b_j,j+1}^{f} \geq t_{ij} & : w_{ij} \\
\dfrac{\sum_{j=1}^{U_M}\sum_{i=D}^{N} e_{ij}^{f}}{N-D} - \epsilon \leq \alpha^* & : \theta
\end{array}
$$

The dual subproblem

$$max\{\sum_{i=2}^{N}\sum_{j=1}^{N_M^r} t_{i,j}u_{i,j} + \sum_{i=1}^{N}\sum_{j=2}^{N_M^r} t_{i,j}v_{i,j} + \sum_{j=1}^{N_M^r-1}\sum_{i=b_j+1}^{N} t_{i,j}u_{i,j} + a_1 t_{1,1} - \alpha^*\theta\}$$

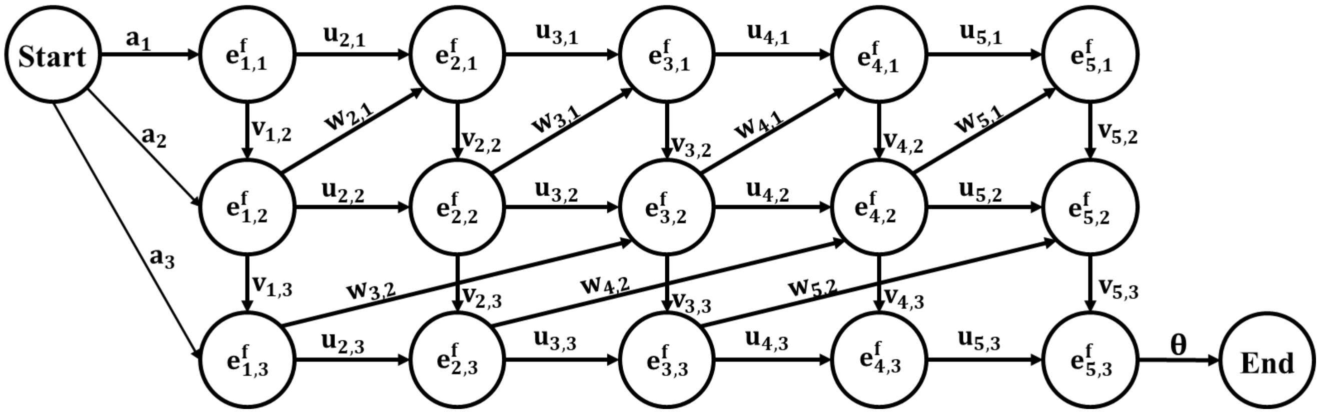
s.t.

$$
\begin{array}{ll}
a_1 - u_{2,1} - v_{1,2} = 1 & \\
v_{1,j} - v_{1,j+1} - u_{2,j} - w_{b_{j-1},j-1} = 1 & 2 \leq j \leq N_M^r - 1 \\
v_{1,N_M^r} - u_{2,N_M^r} - w_{1+b_{N_M^r-1},N_M^r-1} = 1 & \\
u_{i,1} - u_{i+1,1} - v_{i,2} = 1 & 2 \leq i \leq b_1 + 1 \\
u_{i,1} + w_{i,1} - u_{i+1,1} - v_{i,2} = 1 & b_1 + 2 \leq i \leq N - 1 \\
u_{N,1} + w_{N,1} - v_{N,2} = 1 & \\
u_{i,j} + v_{i,j} - u_{i+1,j} - v_{i,j+1} - w_{i+b_{j-1},j-1} = 1 & 2 \leq j \leq N_M^r - 1, 2 \leq i \leq b_j + 1 \\
u_{i,j} + v_{i,j} + w_{i,j} - u_{i+1,j} - v_{i,j+1} - w_{i+b_{j-1},j-1} = 1 & 2 \leq j \leq N_M^r - 1, b_j + 2 \leq i \leq N - b_{j-1} \\
u_{i,j} + v_{i,j} + w_{i,j} - u_{i+1,j} - v_{i,j+1} = 1 & 2 \leq j \leq N_M^r - 1, \\
 & N - b_{j-1} + 1 \leq i \leq N - 1 \\
u_{N,j} + v_{N,j} + w_{N,j} - v_{N,j+1} = 1 & 2 \leq j \leq N_M^r - 1 \\
v_{1,N_M^r} - u_{2,N_M^r} - w_{1+b_{N_M^r-1},N_M^r-1} = 1 & \\
u_{i,N_M^r} + v_{i,N_M^r} - u_{i+1,N_M^r} - w_{i+b_{N_M^r-1},N_M^r-1} = 1 & 2 \leq i \leq N - b_{j-1} \\
u_{i,N_M^r} + v_{i,N_M^r} - u_{i+1,N_M^r} = 1 & N - b_j + 1 \leq i \leq N - 1 \\
u_{N,N_M^r} + v_{N,N_M^r} - \dfrac{\theta}{N} = 1 &
\end{array}
$$

Original contribution: the optimal solution of the dual subproblem can be calculated from simulation.

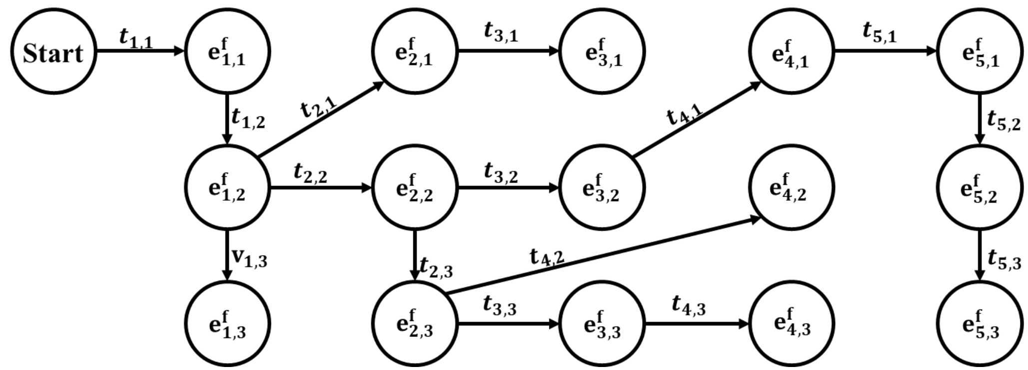
Network flow: Dual Subproblem

- The graph of the network flow problem (the dual subproblem) is the same as the ERG.
- The variables of the dual subproblem are the flows of all arcs.
- Each node e_{ij}^{f} is a sink which absorbs one unit flow.

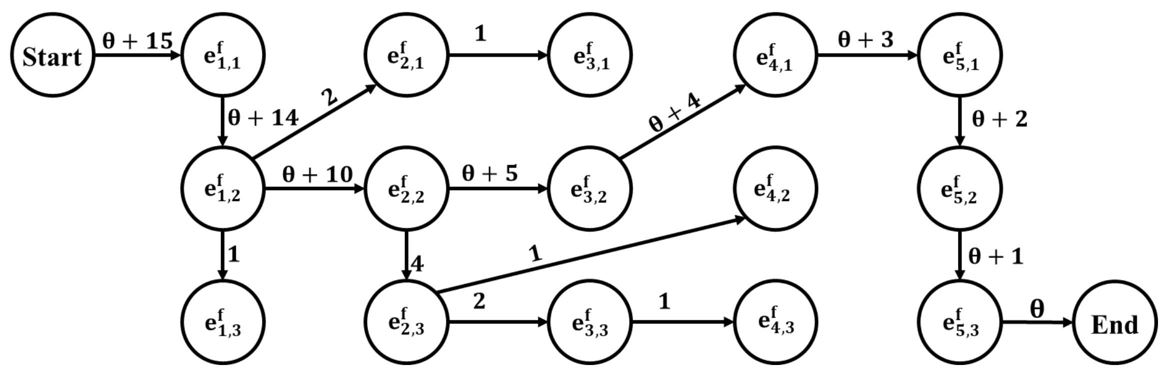


Network flow: Dual Subproblem

- After simulation (with any tool), the ERG becomes a **simulated ERG**.

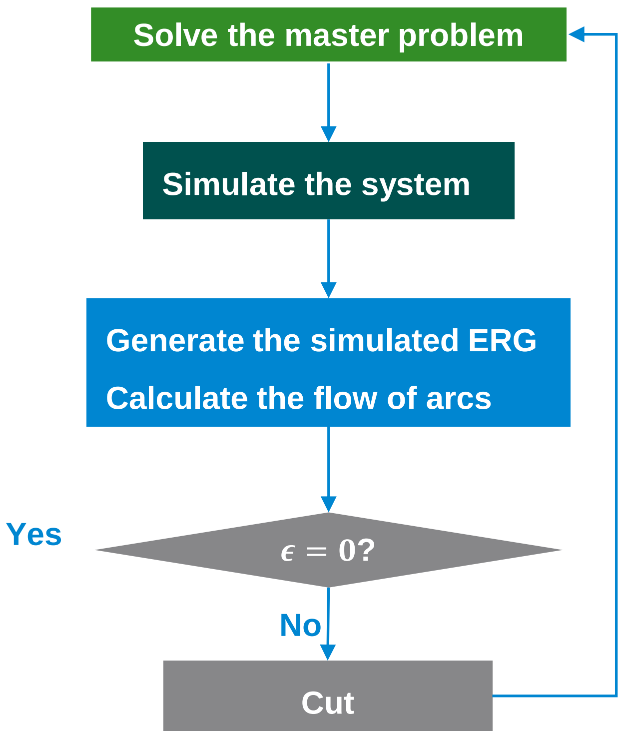


- The optimal solution of the dual subproblem can be derived from the simulated ERG.



- $\boldsymbol{\theta = N_\epsilon}$. (Optimality can be proved)

Simulation cutting approach



- Only the master problem is solved by optimization solvers (e.g., Cplex).
- Simulation is used to solve the network flow problem.
- The cut reflects the simulation event relationship.

FEASIBILITY CUT

$$-M\sum_{j=1}^{U_M-1}\sum_{k=1}^{U_B}\sum_{i=1+b_j}^{N}\bar{x}_{jk}\bar{w}_{jk}\left(1-x_{jk}\right)+\sum_{j=1}^{U_M}\sum_{i=2}^{N}\phi\left(s_j,z_{ij}\right)\left(\bar{u}_{ij}+\bar{v}_{ij}+\bar{w}_{ij}\right)+\bar{a}_1\phi\left(s_j,z_{ij}\right)-\alpha^*\bar{\theta}\le 0$$

Numerical experiments

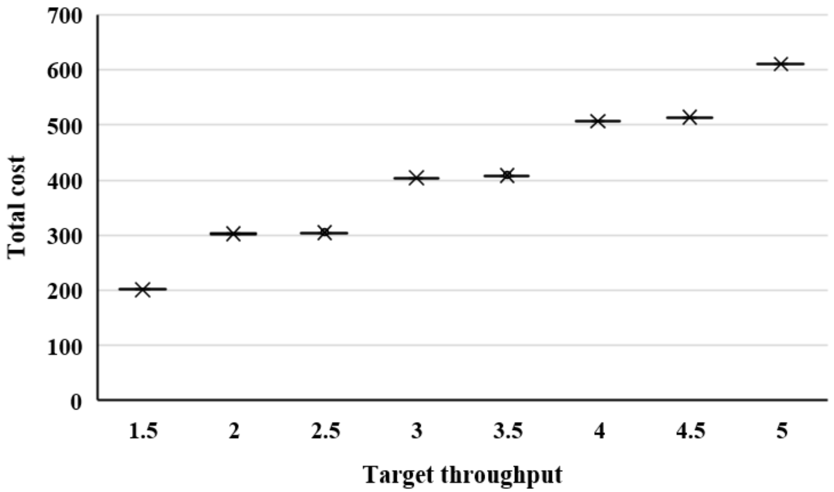
Processing time distribution of all workstations: Beta(2,2)

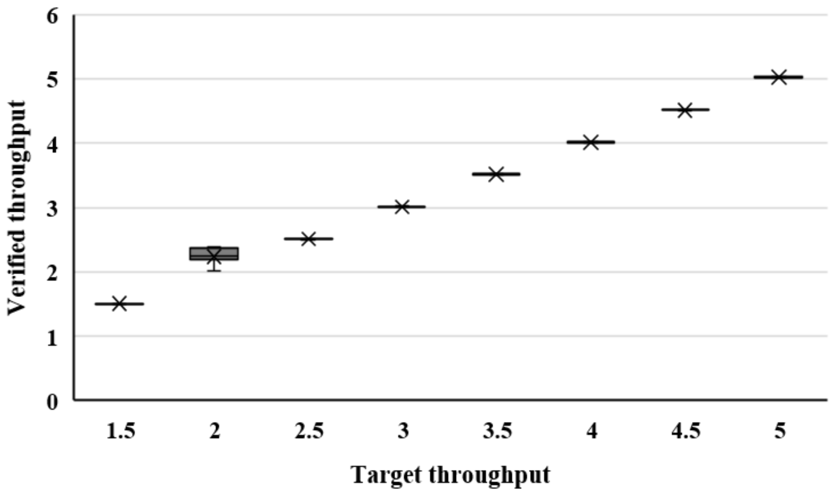
Average total processing time: 1 time unit

Target throughput: 1.5-5 parts/time unit

Number of parts for solution: 100 000

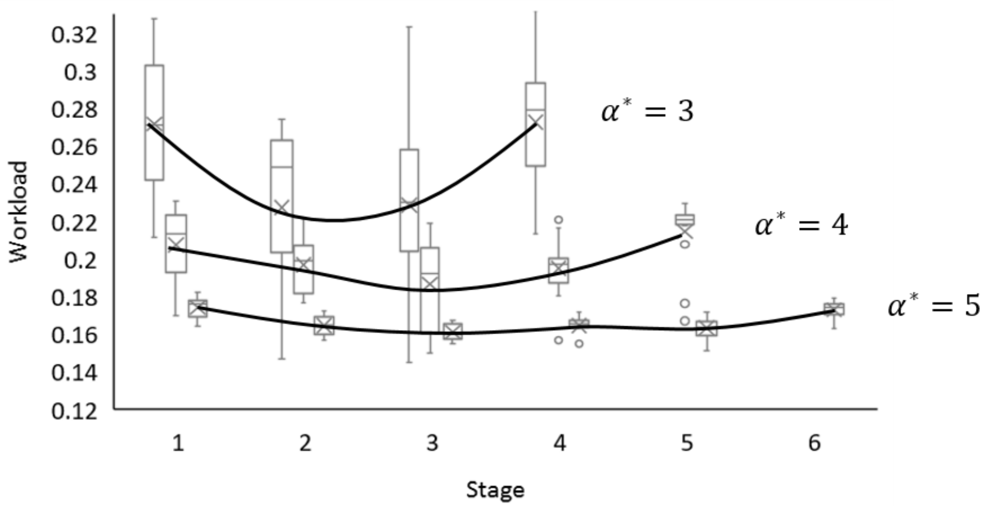
Number of parts for verification: 1000 000





The two graphs are box plots from 10 different sample paths.

Solution Pattern



α^*	Buffer Stage 1	Buffer Stage 2	Buffer Stage 3	Buffer Stage 4	Buffer Stage 5
2	1.1	1.1			
2.5	2	2.3			
3	1.4	1.4	1.2		
3.5	3.1	2.5	2.5		
4	1.6	2	1.5	1.9	
4.5	3.5	3.3	3.2	4	
5	2.0	2.0	2.0	2.0	2.0

The graph and the data are from 10 different sample paths.

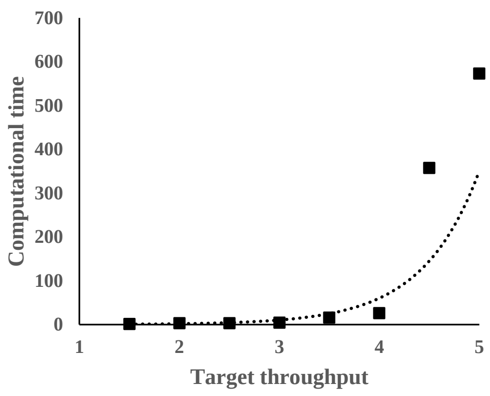
Processing time distribution: Beta(2,2)

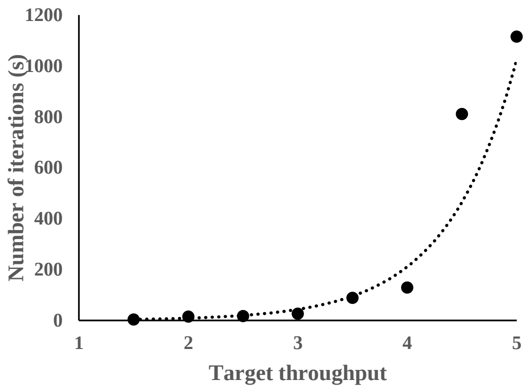
Average total processing time: 1 time unit

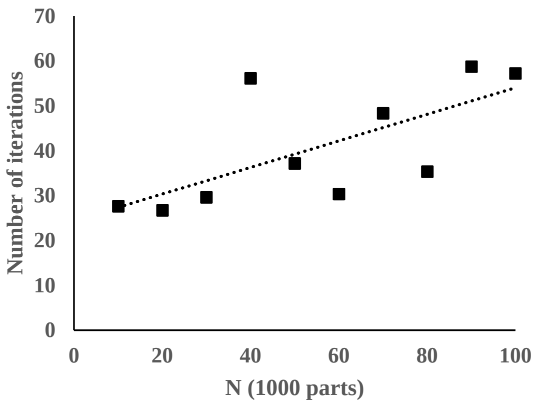
Target throughput: 3, 4, 5 parts/time unit

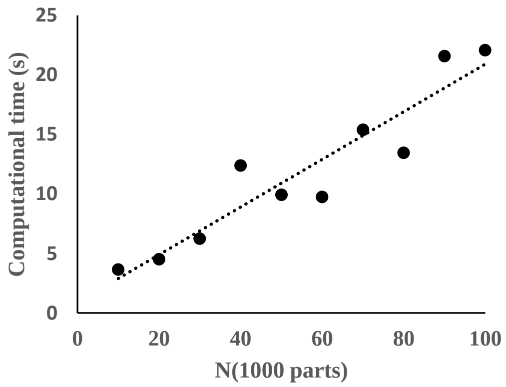
Number of parts for solution: 100 000

Efficiency analysis



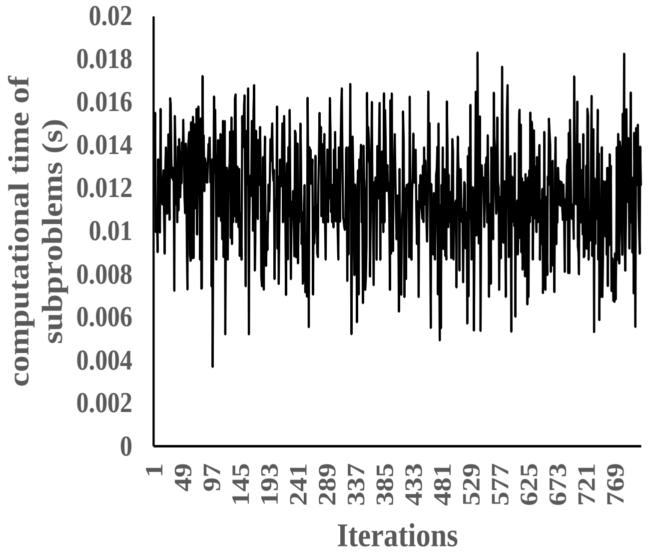




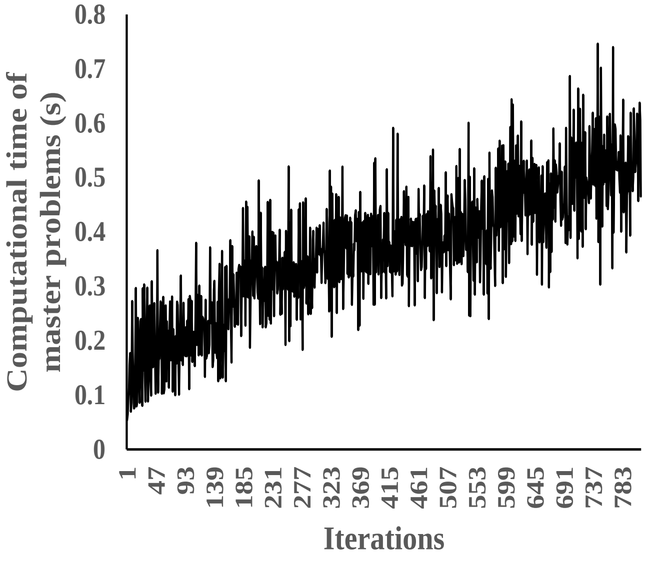


Efficiency analysis

SubProblem



Master Problem



Contribution

- The DEO model of the JWWBAP is **exactly** solved, and the solution is the global optimal based on one sample path.
- The simulation cutting approach uses the **event relationships** in simulation to build the cut.
- Simulation is used as a **white box**: simulation does not only evaluate the optimization output, but recognizes the events which impacts the performance most significantly as well.

Future research

- The simulation cutting approach will be applied to solve more complex DEO models, e.g., G/G/m. In a DEO model of G/G/m system, the subproblem (simulation) is a mixed integer programming model, so the dual subproblem cannot be easily generated.
- As the solution of the master problem takes most of the computational effort, more efficient algorithms for solving the master problem will bring significant improvement of the simulation cutting approach.
- How to manage cut from several independent replications ?

Thank you