Multi-Level Monte Carlo Analysis of manufacturing systems

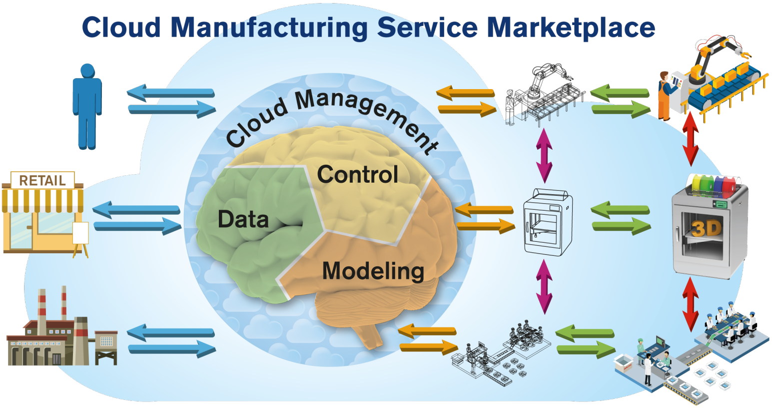
Giulia Pedrielli(*),
Feng Ju(*)

(*) School of Computing Informatics & Decision Systems Engineering,
Arizona State University, giulia.pedrielli@asu.edu

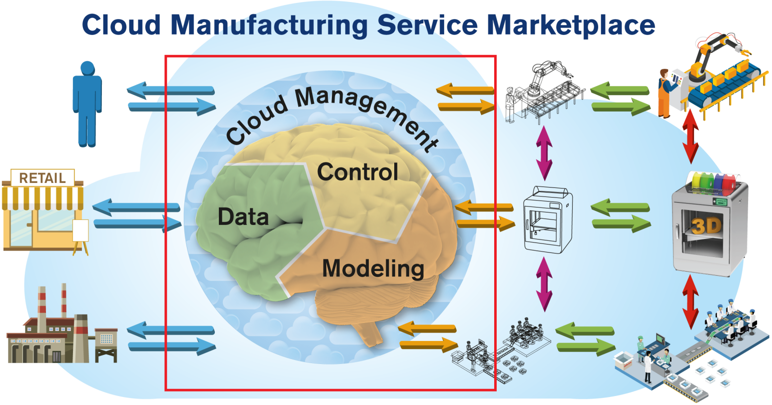
June 5th, 2017

Outline

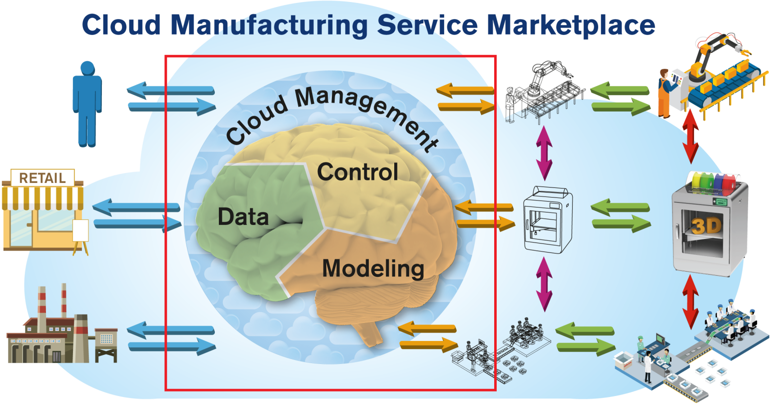
Connected plants and smart systems



Connected plants and smart systems

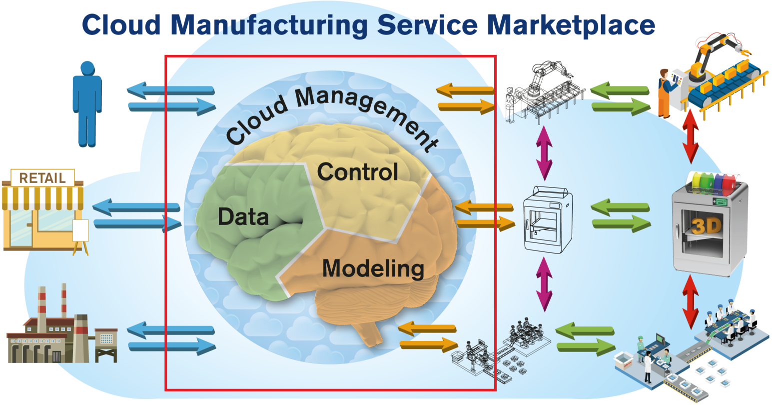


Connected plants and smart systems



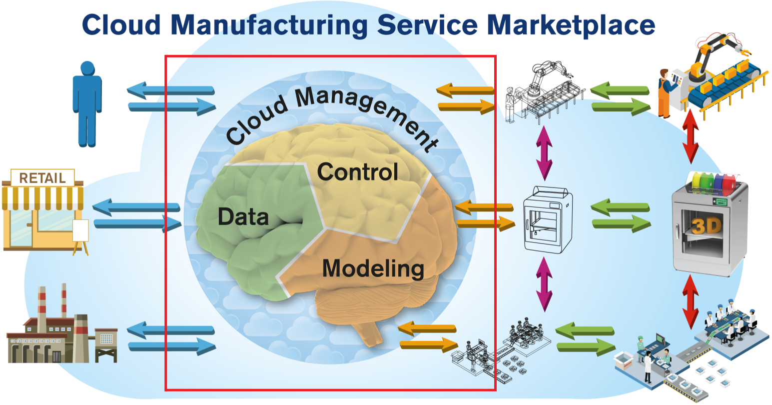
- Multiple and complex parts;

Connected plants and smart systems



- Multiple and complex parts;
- Tasks can be distributed among several connected plants;

Connected plants and smart systems



- Multiple and complex parts;
- Tasks can be distributed among several connected plants;
- New technologies make sharing information easier.

Smart Brain with Advanced Analytics

- **Opportunities and challenges**
  - Global sensors;
  - Real-time dynamic data;
  - Connection and integration.
- **Big Data**
  - Data Analytics is needed for what if;
  - Big Data to support complex decision making.
- **Complexity of interconnected systems**
  - Most decision/OR tools require restrictive assumptions/approximations;
  - While simulation can handle complex systems, computation efficiency is still a concern.

Current Paradigms & Proposed Approach

Category	Deterministic Optimization	Stationary Approximation	Stochastic Optimization	Simulation Optimization	Heuristics
Solution Methods	Mathematical programs	Iterative linear optimization procedures with embedded simulations	Stochastic programs or chance-constrained optimization	Nonlinear direct search using high-fidelity simulations	Rule-based or a hybrid of optimization, simulations, and rules
Non-linearity	N	Y	N	Y	Limited
Uncertainty	N	N	Y	Y	Limited
Transiency	N	N	N	Y	Limited
Computing time	Fast	Fast, but unpredictable	Slow	Very slow	Fast
Solution quality	Poor	Good	Good	Best	Unpredictable

Basic Idea: We need to make simulation optimization quicker.

- Fast performance estimation → Multi-Fidelity Estimation.
- Effective sampling techniques → Ordinal Transformation and Mixed Model Sampling

Performance evaluation for manufacturing systems

Analytical Models

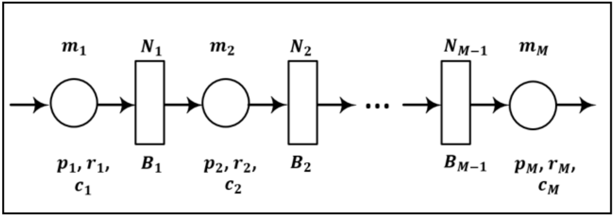
- Queuing models (Perros,1989; Onvural,1990)
- Flow models (Tan, 2013;Levantesi, Matta, Tolio, 2003)
- Markov chain models (Dallery and Gershwin, 1992; Li and Meerkov, 2003)

Simulation Models

- Extreme Flexibility (Law and Kelton 2011);
- Strong statistical support for analysis of output (Glynn et al. 2016);
- Discrete Event and System dynamics are among the most common (Banks et al. 2010).

- Analytical models are usually fast to execute;
- AM link input with output "explicitly";
- simulation models do not require assumptions.

System Description



Notations	
m	Machine index
B	Buffer index
c	Machine speed
p	Failure rate
r	Repairing rate
N	Buffer capacity

Remarks:

1. All machines operate/fail independently.
2. Block before service.
3. There exists a relationship between the processing rate ($c_k = 1/\tau_k$) and failure rate (p_k): for machine k, $p_k = f(c_k)$, where f is a piecewise linear function with ℓ pieces and $\alpha_{k,\ell}$ as proportionality constant.

Problem Formulation

Objective

- Develop a model to evaluate the production rate of the given serial production line with:
  high accuracy and **low computational cost**.
- We want a fast estimation procedure on production rate for perspective system control.

In order to achieve our objective we want to use both **analytical** as well as **simulation** models.

Analytical Model

Two-machine-one-buffer lines

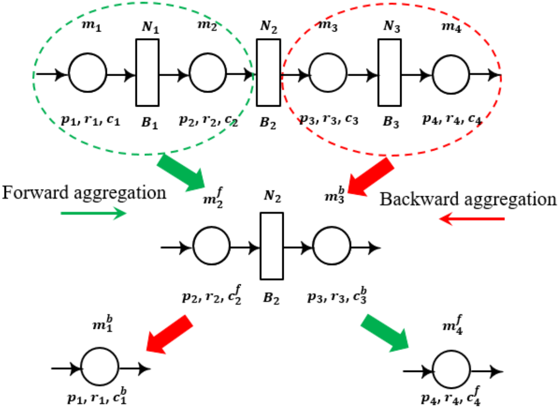
- Building block
  - Modeling into a continuous time discrete state Markov chain, with states defined as a combination of machine status and buffer level.
  - A closed-form equation is obtainable to express the system production rate in terms of system parameters.

$$\widehat{\theta} = F(c_1, c_2, p_1, p_2, r_1, r_2)$$

- The analytical model assumes a simple linear relationship between the processing and failure rates $p_k = 1/\alpha_k(c_k)$.

Analytical Model

- General serial production lines
  No closed-form formula and state aggregation approximation approach is applied.



Creating simulation Models at several fidelities

In principle there are several ways we can control the precision (fidelity) of the simulation model against the computational effort:

- We assume increasing computational time implies increased model precision ($\mathcal{A}_1$);
- An high fidelity simulator is available able to return the true response ($\mathcal{A}_2$);
- Aggregation and disaggregation can be adopted to reduce the complexity of the DES model;
- Given the DES model, we can reduce the simulation run length and/or the number of simulation replications.

We developed an Arena DES and decided to adopt the run length approach to regulate the fidelity of the simulator.

Multi-Level Estimation

When multiple-fidelities are available we are given a chance to increase the precision of the estimator by using *all* the available models.

- Use Gaussian processes to predict unsampled points;
- Use control variates framework to construct an unbiased estimator of the response;
- For every model, compute the optimal weight **given the sampled locations**.

Assuming Gaussian processes for the responses, we can have an analytical form for the MSE, this is essential to assign optimal weights to the models.

Estimation with Gaussian Processes

Each model produces a response of the type:

$$\theta(\boldsymbol{x}) = \theta_k(\boldsymbol{x}) + B_k(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$$

We model $B_k(\boldsymbol{x}) \sim GP\left(\mu_k(\boldsymbol{x}), \tau^2_{B_k}\boldsymbol{R}_{B_k}\right)$.
The proposed estimator will be of the form:

$$\hat{\theta}_{MF}(\boldsymbol{x}) = \sum_{k=1}^{J} w_{k,LF}\hat{\theta}_k(\boldsymbol{x})$$

The weights will be computed as to:

$$\boldsymbol{w}^* \in \arg\min Var\left(\hat{\theta}(\boldsymbol{x})\right)$$

Estimation with Gaussian Processes (cont'd)

- $\mathcal{M}_{1,2}$ considers only the results from the analytical model and uses the bias $B^{LF,a(s)}(\boldsymbol{x})\,|\boldsymbol{x}$ of sample points $\boldsymbol{x}$:

$$\hat{\theta}^{HF}(\boldsymbol{x}) = \theta^{LF,a(s)}(\boldsymbol{x}) + B^{LF,a(s)}(\boldsymbol{x});$$

- $\mathcal{M}_3$ uses a weighted average of the two gaussian processes $B^{LF,a}(\boldsymbol{x})\,|\boldsymbol{x}, B^{LF,s}(\boldsymbol{x})\,|\boldsymbol{x}$ with $w_a = \frac{1}{2}$: $\hat{\theta}^{HF}(\boldsymbol{x}) = w_a\left(\theta^{LF,a}(\boldsymbol{x}) + B^{LF,a}(\boldsymbol{x})\right) + (1-w_a)\left(\theta^{LF,s}(\boldsymbol{x}) + B^{LF,s}(\boldsymbol{x})\right)$.
- $\mathcal{M}_4$ uses the same principle as $\mathcal{M}_3$, but a enhanced estimate through control variates:

$$\hat{\theta}^{HF}(\boldsymbol{x}) = \quad \bar{\theta}^{HF}(\boldsymbol{x}) + \quad \beta^*(\boldsymbol{x}) \begin{bmatrix} \hat{\theta}^{LF,s}(\boldsymbol{x}) - E\left[\hat{\theta}^{LF,s}(\boldsymbol{x})\right] \\ \hat{\theta}^{LF,a}(\boldsymbol{x}) - E\left[\hat{\theta}^{LF,a}(\boldsymbol{x})\right] \end{bmatrix}$$

Multi-level Estimation: control variates

- $\hat{\theta}_k^{LF}$ is the estimation of the high fidelity response returned from the low fidelity model. We need a very low number of high fidelity simulation n_0 to estimate this;
- Conditional on the location $\boldsymbol{x}$, we consider $\vartheta_k^{LF}|\boldsymbol{x} \sim \mathcal{N}\left(\mu_k^{LF}, \left[\sigma_k^{LF}\right]^2\right)$;
- We can generate a large number of low fidelity estimates of the high fidelity model with $n_0, N^{HF} > n_0, \Delta_k^{LF} = (1+\gamma_k)\, N^{HF}$:
  - n_0 is used to create the low fidelity generator;
  - N^{HF} is used to generate the high-fidelity estimate;
  - Δ_k^{LF} are bootstrapped observations from the low fidelity generator.

Multi-level Estimation: control variates

- MLMF estimator:

$$\hat{\theta} = \hat{\theta}^{HF} + \beta^T \begin{bmatrix} \hat{\theta}_1^{LF} - E\left[\hat{\theta}_1^{LF}\right] \\ \hat{\theta}_2^{LF} - E\left[\hat{\theta}_2^{LF}\right] \end{bmatrix}$$

- Variance of the estimator:

$$Var\left(\hat{\theta}\right) = \frac{1}{N^{HF}} Var\left(\hat{\theta}^{HF}\right) + \sum_k \left[\frac{\beta_k^2 \gamma_k}{(1+\gamma_k)(N^{HF}-n_0)} Var(\vartheta_k^{LF})\right]$$
$$+2\sum_k \left[\frac{\beta_k \gamma_k}{(1+\gamma_k)(N^{HF}-n_0)} \rho_{HL} \sqrt{\left(Var\left(\hat{\theta}^{HF}\right) Var(\vartheta_k^{LF})\right)}\right]$$

- Minimum variance weight:

$$\beta_k^* = -\rho_{HL} \frac{\sqrt{Var\left(\hat{\theta}^{HF}\right)}}{\sqrt{Var(\vartheta_k^{LF})}}$$

Experimental Settings

1. For $\mathcal{HF}$, replication number$=50$, length per replication $=5000$; for $\mathcal{M}_2$, replication number $=5$, length per replication $=500$, 750, and 1000 for three experiments;
2. Set the parameters for the system, $M=3, N_1=3, N_2=2, R_1=0.55, R_2=0.56, R_3=0.55$; $P_1=0.1+0.04c_1,\ P_2=0.15+0.03c_2,\ P_3=0.1+0.05c_3$.
3. Generate 50 sets of capacity speed triples, with speed in range $(0.5,1)$. Predict for model $\mathcal{M}_1, \mathcal{M}_2$, and $\mathcal{HF}$.
4. Calculate the bias and train model $\mathcal{M}_3$ and $\mathcal{M}_4$, and then generate $10,000$ prediction points from each model.
5. Randomly select $3,000$ points from the $10,000$ prediction points, run high fidelity simulations and measure accuracy with:

$$\delta_{model} = \frac{|\widehat{\theta}^{\mathcal{M}} - \theta^{HF}|}{\theta^{HF}} \times 100\%$$

Results & Discussion

Table 1: Summary for the performance of different prediction models

δ_{model}	length=500		length=750		length=1000	
	Mean	STD	Mean	STD	Mean	STD
$\mathcal{M}_1$	0.2095	0.0952	0.2095	0.0952	0.2095	0.0952
$\mathcal{M}_2$	0.1992	0.1217	0.2024	0.1257	0.2034	0.1231
$\mathcal{M}_3$	0.2506	1.4226	0.2585	1.4465	0.2446	0.6569
$\mathcal{M}_4$	0.1117	0.0825	0.1101	0.0818	0.1125	0.0833
$\mathcal{HF}$	0.2075	0.1409	0.2075	0.1409	0.2075	0.1409

1. The combined bias model $\mathcal{M}_4$ yields the best result.
2. There is no significant difference when increasing the per replication length of the low fidelity simulation.
3. The uniform weight model $\mathcal{M}_3$ does not provide satisfactory results!

Conclusions & Future Work

Conclusions

- It is possible to improve the quality of the estimations combining models of different fidelities;
- Gaussian Processes show a good performance even with the naive estimator;
- Weighting the models has to be performed in a clever way and uniform weighting can lead to poor results.

Future Work

- Extend to the optimization phase;
- Explore different models from Gaussian Process;
- Sampling in different fidelities means to solve a dynamic allocation process (γ_k^*).

Thank You

Dr. Giulia Pedrielli,
giulia.pedrielli@asu.edu
Dr. Feng Ju, fengju@asu.edu .