Efficient Simulation Sampling Allocation Using Multi-Fidelity Models

Jie Xu

Dept. of Systems Engineering & Operations Research

George Mason University

Fairfax, VA

jxu13@gmu.edu

Joint work with Y. Peng, C.-H. Chen, L.-H. Lee, J.-Q. Hu

Supported by NSF and AFOSR under Grants ECCS-1462409 and CMMI-1462787

SIMULATION-BASED DECISION MAKING

- **Simulation provides a predictive tool for decision making when problems are intractable to analytical approaches**
- **This talk considers a special case known as ranking & selection**
  - $x_{[1]} = \text{argmax}_{i \in \{1,2,\ldots,I\}} f(x_i)$
  - Stochastic black-box objective functions, observed by running iid replications of a simulation model
- **Fruitful research on simulation-based decision making**
  - Efficient sampling/allocation of simulation budget, convergent fast local search, parallelization, surrogate model
  - Open-source solver ISC (www.iscompass.net) has been used by MITRE and the Argonne National Lab in real-world problems air traffic management and power systems applications
  - What if the full-scale simulation model runs for hours?

CAN APPROXIMATION MODELS HELP?

Full-featured model	Approximation model
High-fidelity/full-scale discrete-event simulation, agent-based model, etc.	Low-fidelity/reduced-scale simulation, analytical approximation, full-model with archived data
Complex	Simple
Accurate	Approximate
Time-consuming	Fast

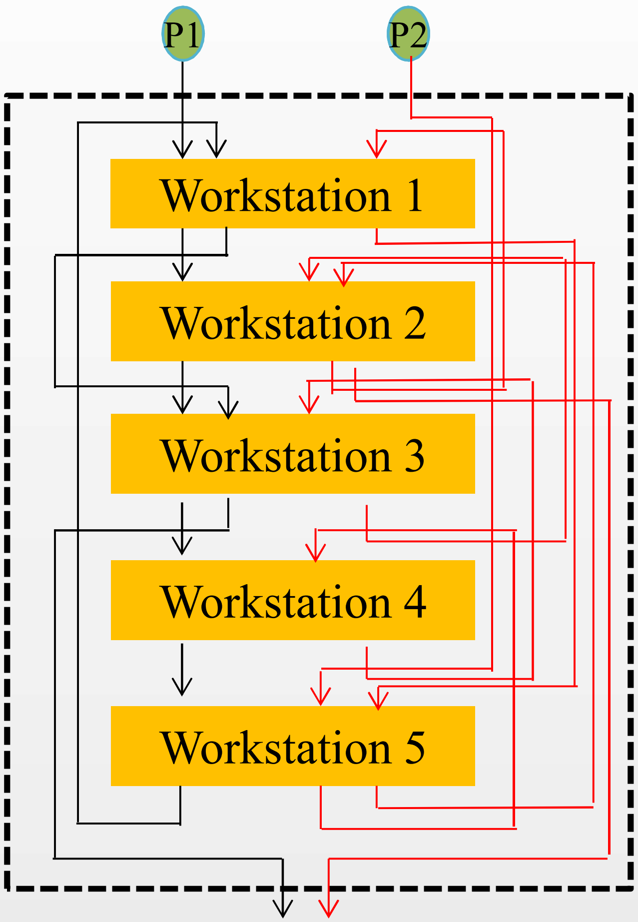
MULTI-FIDELITY OPTIMIZATION METHODS

- **A naïve way of multi-fidelity optimization**
  - Find some most promising designs using the approximation model
  - Evaluation using high-fidelity simulations
- **Most approaches use interpolation/regression to "correct" low-fidelity model**
  - Autoregressive framework with kriging/Gaussian process regression (Kennedy and O'Hagan 2000)
  - Radial basis function, Polynomial chaos
- **Significant challenges arise when**
  - Solution space is high-dimensional
  - High-fidelity simulation samples have heterogeneous noise
  - Quality of low-fidelity model is low
  - Mixed decision variables (integer, categorical)

Simulation Optimization: An Illustrative Example

Resource allocation problem in a flexible manufacturing system

- 2 product types
- 5 workstations
- Non-exponential service times
- Re-entrant manufacturing process
- Product 1 has higher priority than product 2



<u>Optimization problem:</u>

Decision variable

Number of machines at each workstation

Objective

Minimize Expected Total Processing Time

EXAMPLE: RESOURCE ALLOCATION PROBLEM

Decision variables: number of machines allocated to each workstation

$\boldsymbol{Minimize}$

$$\text{Total Processing Time}$$

$\boldsymbol{Subject\ to}$

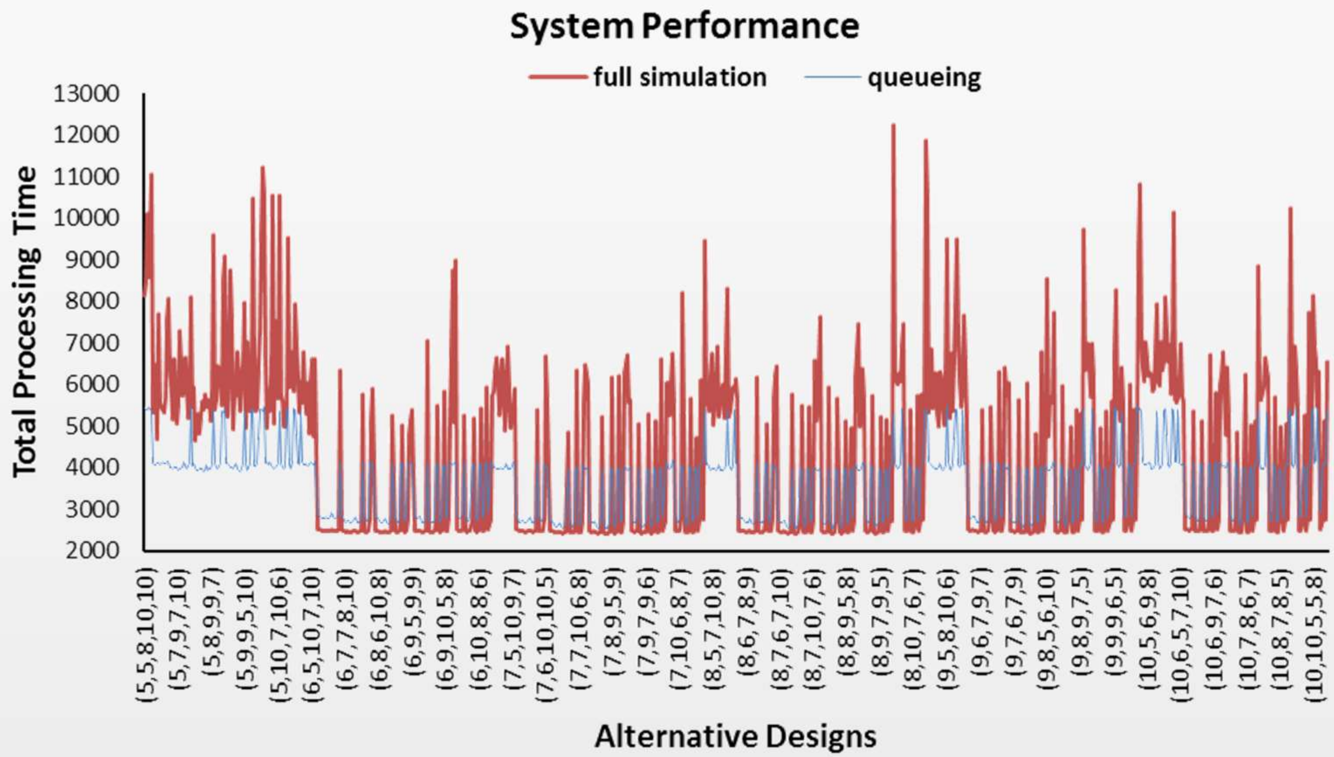
$$\mathbf{5} \leq \#\text{ of machines at each workstation} \leq \mathbf{10}$$
$$\text{Total }\#\text{ of machines at all workstatiosn} = \mathbf{38}$$

$\#$ of alternatives: **780**

→ **Simulation/evaluation can be time consuming**
→ **Solution space dimension can be large**

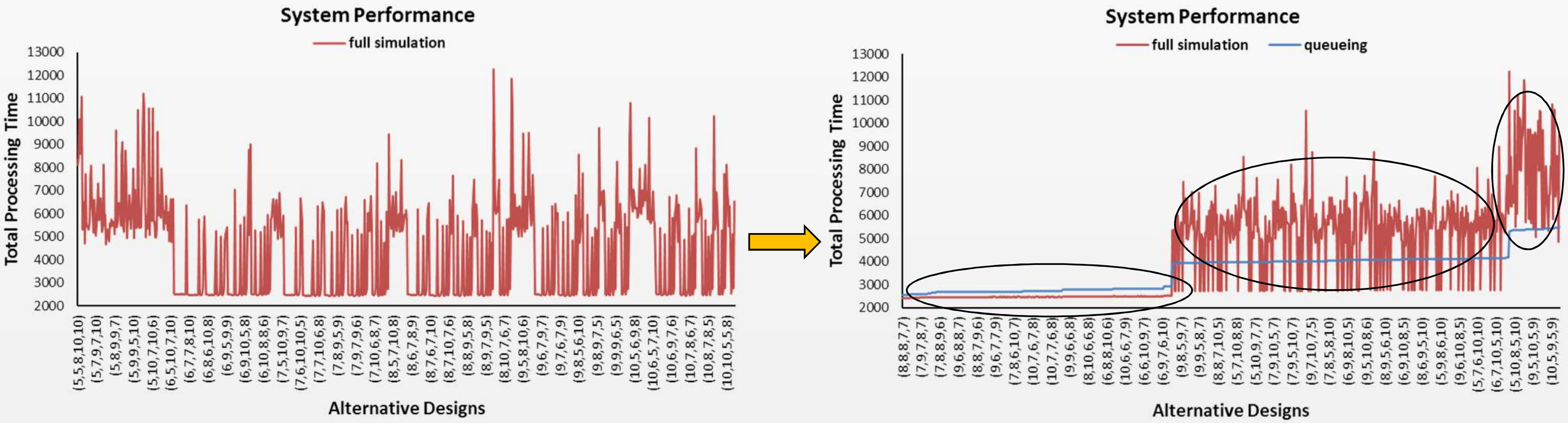
FULL SIMULATION & APPROXIMATION MODELS

Approximation using M/M/c equations, $\rho = 0.83$



Bias is non-homogeneous and can be quite large

ORDINAL RANKINGS OF DESIGNS BY LOW-FIDELITY MODEL



Designs with similar performance are grouped together, which may potentially enhance search/optimization efficiency

A BAYESIAN FRAMEWORK FOR MULTI-FIDELITY MODELS

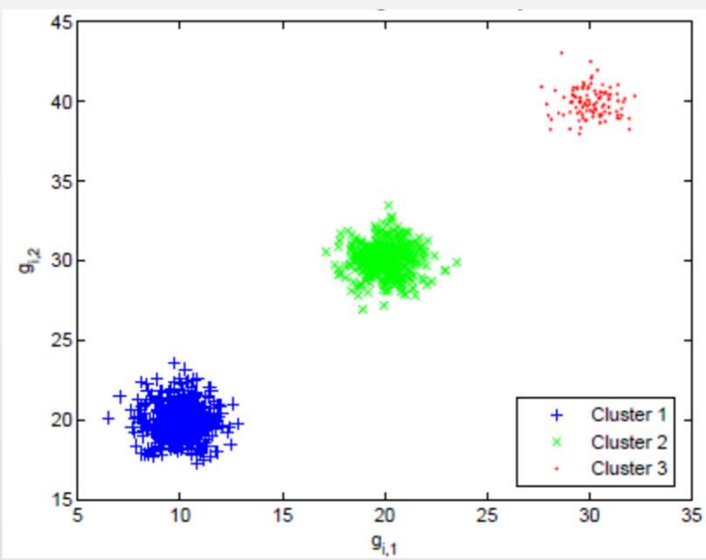
- **For design i, $i = 1, 2, \ldots, k$, we model the prior distribution of high-fidelity (f) prediction and *d*-dimensional low-fidelity predictions ($\vec{g}$) by a Gaussian mixture model**
  - $\vec{h}_i = (f_i, \vec{g}_i) \sim \sum_{m=1}^{M} \tau_m \phi(\cdot \,|\vec{\mu}_m, \Sigma_m)$
  - $\vec{\mu}_m = (\alpha_m, \vec{\beta}_m), \Sigma_m = \begin{bmatrix} \eta_m & \Gamma_m \\ \Gamma_m^T & \Lambda_m \end{bmatrix}, \Sigma_m^{-1} = \begin{bmatrix} \upsilon_m & \Upsilon_m \\ \Upsilon_m^T & \Omega_m \end{bmatrix}$
- **f_i can only be observed with a Gaussian noise $N(0, \sigma_i^2)$**
- **$\vec{g}_i$ is completely observed (negligible computing cost)**
  - $G = (\vec{g}_1, \vec{g}_2, \ldots, \vec{g}_k)$
- **We allocate a total of N high-fidelity simulation replications to designs**
  - Let D_n denote the samples collected after n simulation replications
  - Let n_i be the number of simulation replications allocated to design i after n simulation replications

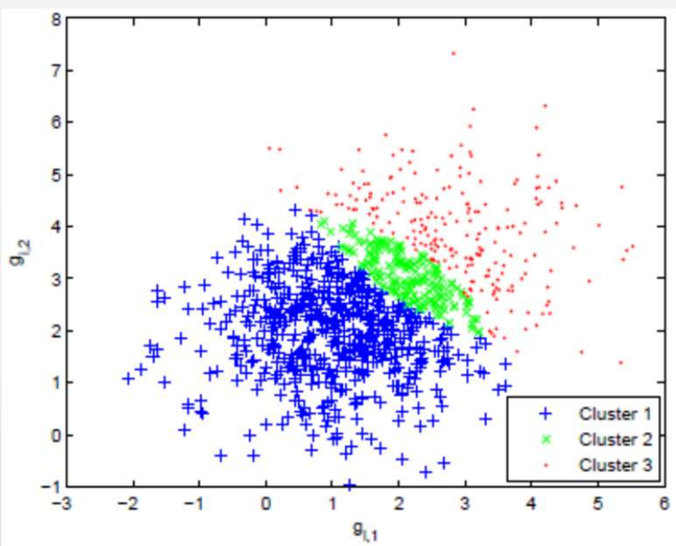
MODEL ESTIMATION

- **We extend classical model-based clustering results to the multi-fidelity setting with stochastic observations of $\boldsymbol{f}$**
  - Binary hidden state random variable $z_{i,m}$ assigns design i to cluster m
  - $z_i = \left(z_{i,1}, \ldots, z_{i,M}\right)$ follows a multinomial distribution with parameters $(\tau_1, \ldots, \tau_M)$
- **The maximal likelihood estimate of model parameters $\boldsymbol{\theta_M} = \{\tau_m, \vec{\mu}_m, \Sigma_m\}_{\boldsymbol{m=1}}^{\boldsymbol{M}}$**
  - $\hat{\theta}_M^{(n)} = \arg \max_{\theta_M \in \Theta_M} \mathcal{L}(D_n, G; \theta_M)$
  - $\mathcal{L}(D_n, G; \theta_M) = \prod_{i=1}^{k} \left[\sum_{m=1}^{M} \tau_m \int_{\mathbb{R}} \prod_{l=1}^{n_i} \phi\left(x_{i,l} | f_i, \sigma_i^2\right) \phi\left(\vec{h}_i | \vec{\mu}_m, \Sigma_m\right) d f_i\right]$
- **The Expectation-maximization (EM) algorithm is applied to compute $\hat{\theta}_M^{(n)}$**

MODEL ESTIMATION-CONT.

- **We estimate the number of components M using the completely observed low-fidelity estimates G**
- **Bayesian information criterion (BIC) is used to select M**
  - $BIC_M = \log \mathcal{L}_g(G; \hat{\xi}_M) - \left[\frac{(d+1)(d+2)}{2} M - 1\right] \frac{\log k}{2}$
  - $\hat{\xi}_M = \arg \max_{\xi_M \in \Xi_M} \mathcal{L}_g(G; \xi_M)$, where $\xi_M = \{\tau_m, \vec{\beta}_m, \Lambda_m\}_{m=1}^{M}$
  - $\mathcal{L}_g(G; \xi_M) = \prod_{i=1}^{k} \left[\sum_{m=1}^{M} \tau_m \phi\left(\vec{g}_i | \vec{\beta}_m, = \Lambda_m\right)\right]$
  - We select the M from a specified interval that has the largest BIC_M





THEOREM 1: STOCHASTIC MODEL-BASED CLUSTERING

- **After EM iteration *t*, the posterior probability of $\{z_{i,m} = 1\}$ conditional on D_n and given $\hat{\theta}_M^{(n,t)}$ is**
  - $\hat{z}_{i,m}^{(n,t)} = \frac{\hat{\tau}_m^{(n,t)} C_{i,m}^{(n,t)}}{\sum_{j=1}^{M} \hat{\tau}_j^{(n,t)} C_{i,j}^{(n,t)}}$, where
  - $C_{i,m}^{(n)} \propto \sqrt{\frac{\upsilon_{i,m}^{(n,t)}}{\left|\Sigma_m^{(n,t)}\right|}} \exp\left\{\frac{1}{2}\left[\frac{\left(f_{i,m}^{(n,t)}\right)^2}{\upsilon_m^{(n,t)}} - \hat{\nu}_m^{(n,t)}\left(\hat{\alpha}_m^{(n,t)}\right)^2 + 2\hat{\alpha}_m^{(n,t)}\left(\vec{g}_i - \hat{\beta}_m^{(n,t)}\right)\left(\hat{\Upsilon}_m^{(n,t)}\right)^T - \left(\vec{g}_i - \hat{\beta}_m^{(n,t)}\right)\hat{\Omega}_m^{(n,t)}\left(\vec{g}_i - \hat{\beta}_m^{(n,t)}\right)^T\right]\right\}$,
  - $\hat{\alpha}_m^{(n,t)} = \frac{\sum_{i=1}^{k} \hat{z}_{i,m}^{(n,t-1)} f_{i,m}^{(n,t-1)}}{\sum_{i=1}^{k} \hat{z}_{i,m}^{(n,t-1)}}$, $\hat{\hat{\beta}}_m^{(n,t)} = \frac{\sum_{i=1}^{k} \hat{z}_{i,m}^{(n,t-1)} \vec{g}_i}{\sum_{i=1}^{k} \hat{z}_{i,m}^{(n,t-1)}}$, $\upsilon_{i,m}^{(n,t)} = \frac{1}{\frac{n_i}{\sigma_i^2} + \hat{\nu}_m^{(n,t)}}$

THEOREM 1: STOCHASTIC MODEL-BASED CLUSTERING

- **The posterior distribution of f_i conditional on $\{z_{i,m} = 1\}$, D_n, G, and given $\hat{\theta}_M^{(n,t)}$ is normal with density function $\phi\left(f_{i,m}^{(n,t)}, \upsilon_{i,m}^{(n,t)}\right)$**
  - $f_{i,m}^{(n,t)} = \upsilon_{i,m}^{(n)} \left[\frac{n_i}{\sigma_i^2} \bar{f}_i^{(n)} + \hat{\upsilon}_m^{(n,t)} \hat{\alpha}_m^{(n,t)} - \left(\vec{g}_i - \hat{\hat{\beta}}_m^{(n,t)} \right) \left(\hat{\Upsilon}_{i,m}^{(n,t)} \right)^T \right]$

Weighted high-fidelity simulation sample mean

Weighted cluster mean

Weighted prediction using low-fidelity predictions

- **The estimates of the model parameters are updated in the next EM iteration accordingly**
- **The above results can be extended for noisy $\vec{g}_i$**

Asymptotic results

- **Corollary 1: Suppose that design i is sampled infinitely often as $n \to \infty$, then**
  - $\lim_{n\to\infty}\left[\frac{C_{i,m}^{(n,t)}}{\sum_{j=1}^{M} C_{i,j}^{(n,t)}} - \frac{\phi\left(\vec{h}_i|\hat{\mu}_m^{(n,t)},\Sigma_m^{(n,t)}\right)}{\sum_{j=1}^{M}\phi\left(\vec{h}_i|\hat{\mu}_j^{(n,t)},\Sigma_j^{(n,t)}\right)}\right] = 0$ almost surely
  - This result is consistent with the classical model-based clustering result with $C_{i,m}^{(n,t)}$ playing the role of $\phi\left(\vec{h}_i|\hat{\mu}_m^{(n,t)},\Sigma_m^{(n,t)}\right)$ when the effect of stochastic simulation noise is eliminated
- **Using asymptotic results, we obtain lightweight approximations for posterior estimates that do not require EM iteration**

Asymptotically Optimal sampling allocation Policy

- **Allocate $W = \{w_1, \dots, w_I\}$ high-fidelity simulations to $x_1, \dots, x_I$ to maximize the large deviation rate of incorrect selection event**
- **The large deviation rate of $P(f_i < f_j)$ when $\bar{f}_i^{(n)} > \bar{f}_j^{(n)}$ is given by**

$$G_{i,j}(w_i, w_j) = \frac{(f_i - f_j)^2}{2\left(\frac{\sigma_i^2}{w_i} + \frac{\sigma_j^2}{w_j}\right)}$$

- **Define an approximate large deviation rate (ALDR)**

$$ALDR(W) \triangleq \min_{i \neq 1_{[n]}} G_{1_{[n]},i}\left(w_{1_{[n]}}, w_i\right), \text{ where } 1_{[n]} \triangleq \underset{i=1,\dots,I}{\operatorname{argmax}} \bar{f}_i^{(n)}$$

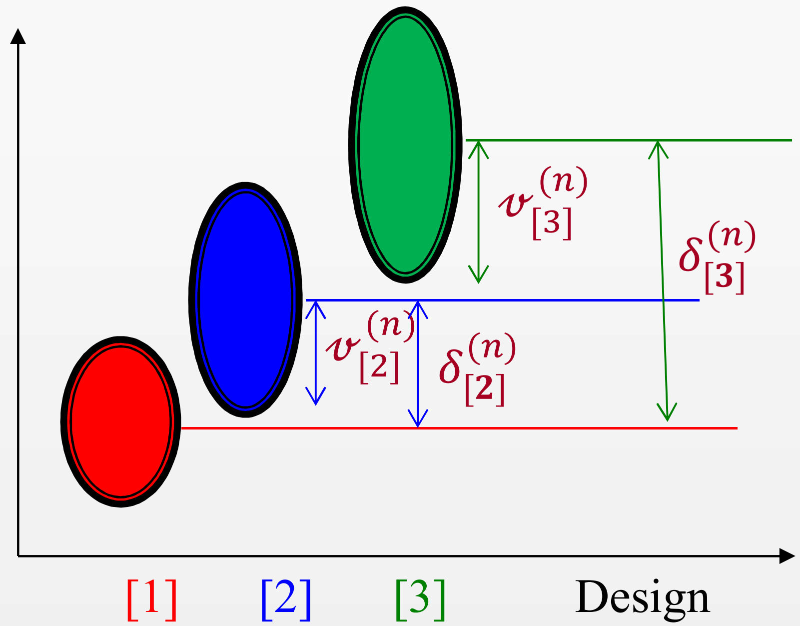
- **It can be shown that $ALDR(W)$ converges with probability 1 to an upper bound on the large deviation rate of incorrect selection event**

Multi-fidelity budget allocation policy

- **Based on the clustering statistics, we define the following posterior means and variances**
  - $f_i^{(n)} = f_{i,\hat{m}_i}^{(n)}, v_i^{(n)} = v_{i,\hat{m}_i}^{(n)}, i = 1, \ldots, k$, where $\hat{m}_i$ is the cluster with the largest clustering statistic for design i
  - Let $[i]$ be the design index after sorting all designs in descending order posterior means, i.e., $f_{[1]}^{(n)} > \cdots > f_{[k]}^{(n)}$
  - Let $\delta_i^{(n)} = \left(f_{[1]}^{(n)} - f_i^{(n)}\right)^2$
- **The (approximately) optimal sampling allocation policy can be obtained by solving**

$$\frac{w_{[i]}}{w_{[j]}} = \frac{v_{[i]}^{(n)} \delta_{[j]}^{(n)}}{v_{[j]}^{(n)} \delta_{[i]}^{(n)}} \text{ for } i, j \neq 1, \;\; w_{[1]} = \sqrt{v_{[1]}^{(n)}} \sqrt{\sum_{i \neq 1} \frac{w_{[i]}^2}{v_{[i]}^{(n)}}}$$

UNDERSTANDING THE SAMPLING ALLOCATION POLICY



Signal to Noise Ratio

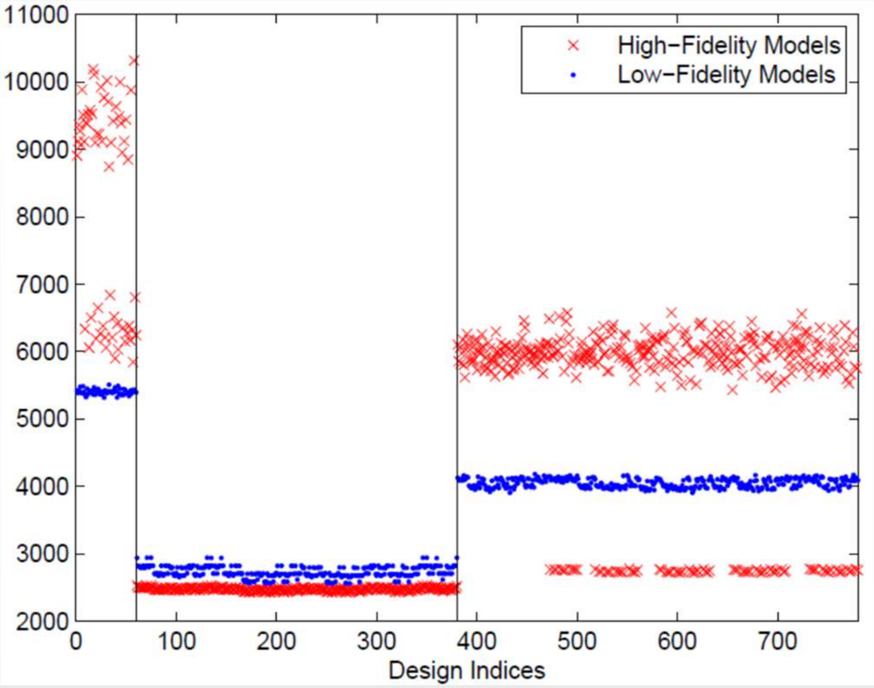
$$\frac{w_{[i]}}{w_{[j]}} = \frac{\delta_{[j]}^{(n)} / v_{[j]}^{(n)}}{\delta_{[i]}^{(n)} / v_{[i]}^{(n)}}$$

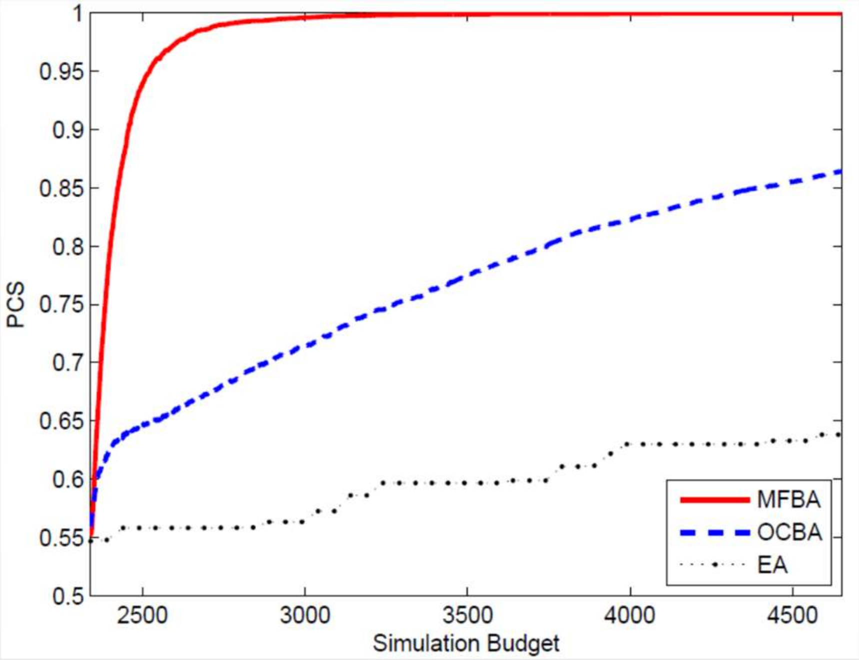
inversely proportional to

the square of the signal to noise ratio

MACHINE ALLOCATION RESULTS

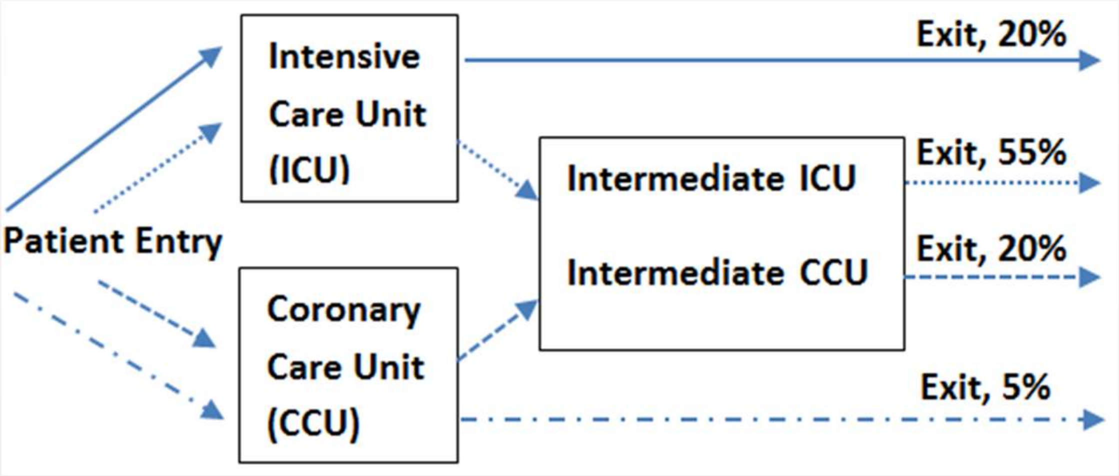
- **Compare the PCS achieved by the new multi-fidelity budget allocation policy (MFBA) with optimal computing budget allocation (OCBA) for one fidelity level and equal allocation (EQ)**





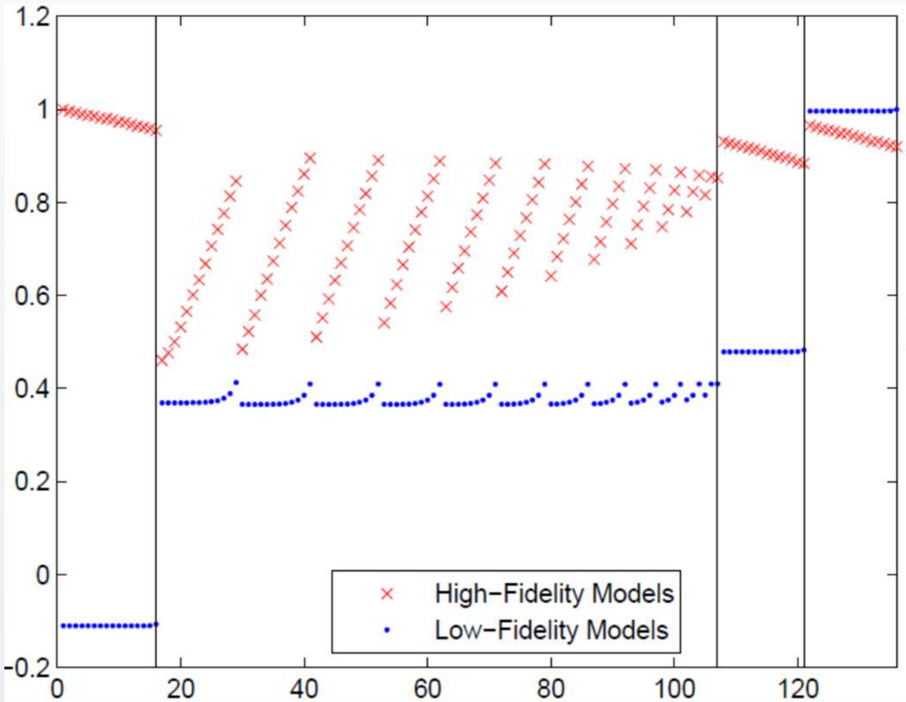
CRITICAL CARE FACILITY RESOURCE ALLOCATION

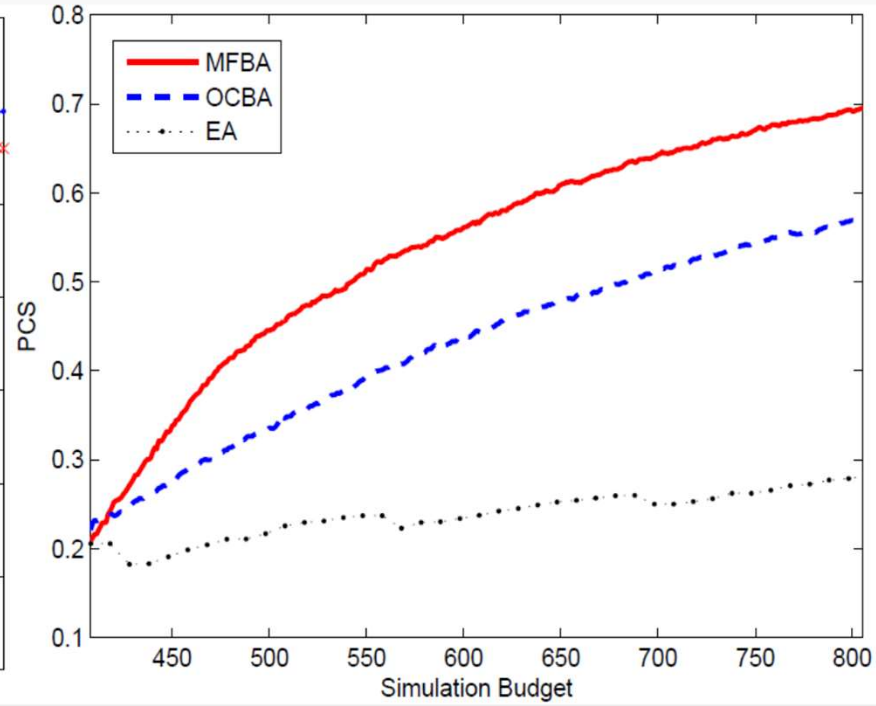
- Allocate 15 additional beds to four care units to reduce the number of patients denied admission because no bed is available at ICU/CCU
- The low-fidelity model is based on M/M/c equations but has poor quality due to limited buffer space and unstable systems



RESULTS

- **Compare the PCS achieved by the new multi-fidelity budget allocation policy (MFBA) with OCBA EQ**





CONCLUSIONS

- **We present a new Bayesian framework with a Gaussian mixture model prior to utilize multi-fidelity information to improve simulation sampling efficiency for the selection of the best design**
- **The multi-fidelity budget allocation policy significantly improves sampling efficiency compared to a single-fidelity optimal sampling policy**
- **Future research includes**
  - Multi-fidelity simulation optimization methods for large-scale problems
  - Incorporation of design co-variates information
  - …
- **Thank you!**