# Modeling for Reliability Optimization of System Design and Maintenance Based on Markov Chain Theory

Yixin Ye[a], Ignacio E. Grossmann[a,*], Jose M. Pinto[b], Sivaraman Ramaswamy[b]

[a] *Department of Chemical Engineering, Carnegie Melon University, Pittsburgh, PA 15232*
[b] *Business and Supply Chain Optimization, Praxair, Inc., Tonawanda, NY 14150*

**Abstract**

This paper proposes an MINLP model that represents the stochastic process of system failures and repairs as a continuous-time Markov chain, based on which it optimizes the selection of redundancy and the frequency of inspection and maintenance tasks for maximum profit. The model explicitly accounts for every possible state of the system. Effective decomposition and scenario reduction methods are also proposed. A small example with two processing stage is solved to demonstrate the impact of incorporating maintenance considerations. A decomposition method and a scenario reduction method are applied to this example and are shown to drastically reduce the computational effort. A larger example with four stages, which is not directly solvable, is also successfully solved using the proposed algorithm. Lastly, we show that the proposed model and algorithms are capable of solving a practical problem based on the air separation process example that motivated our work, which features multiple stages, potential units and failure modes.

*Keywords:* reliability design, maintenance, optimization, Markov Chain, MINLP

## 1. Introduction

Plant availability has been a critical consideration for the design and operation of chemical processes as it represents the expected fraction of normal operating time, which directly impacts the ability of making profits. In practice, discrete-event simulation tools are used to examine the availability of a few selected designs of different redundancy levels under various maintenance and spare parts inventory policies (Sharda and Bury, 2008). However, the best plan selected through simulation is usually suboptimal because the list of design alternatives is often not exhaustive. Thus, there is a strong motivation for systematic optimization tools of redundancy design considering operational factors.

Several works have been reported regarding reliability considerations at the design phase. Kuo and Wan (2007) give a literature survey on optimal reliability design methods classified in terms of problem formulations and optimization techniques. Thomaidis and Pistikopoulos (1994, 1995) integrate flexibility and reliability in process design without considering the possibility of having standby units. Aguilar et al. (2008) address the reliability issue in utility plant design and operation by considering some pre-specified alternatives for redundancy, and for which they formulate an

*corresponding author
*Email address:* grossmann@cmu.edu (Ignacio E. Grossmann)

MILP model considering a few number of failure scenarios. Ye et al. (2017) propose a general mixed-integer framework for the optimal selection of redundant units bearing reliability concerns. Jensen et al. (2016) develop a model-reduction strategy for reliability-based design problems of complex structural systems to reduce the computational efforts of solving finite element models.

In order to obtain a more comprehensive optimal design, it is important to consider the impact of operational factors on plant availability and their costs. For example, maintenance, is a major strategy to improve the availability of units (Ding and Kamaruddin, 2015). Many works have been reported on the modeling and optimization of maintenance (Sharma et al., 2011). Alaswad and Xiang (2017) provide a review for condition-based maintenance optimization models for stochastically deteriorating system with either discrete or continuous states. Chiang and Yuan (2001) propose a state-dependent maintenance policy for a multi-state continuous-time Markovian deteriorating system. Lee and Cha (2016) describe a preventive maintenance optimization model under the assumption that the failure process follows a generalized Poisson process. Pistikopoulos et al. (2001) and Goel et al. (2003) formulate MILP models for the selection of units with different reliability and the corresponding production and maintenance planning for a fixed system configuration. Also, these works optimize the maintenance schedule balancing maintenance costs and the benefits from availability increase (Vassilliadis and Pistikopoulos (1999), Cheung et al. (2004), Nguyen and Bagajewicz (2008)).

Markov chain is a powerful mathematical tool that is extensively used to capture the stochastic process of systems transitioning among different states. Shin and Lee (2016) formulate the planning level problem of a procurement system as an Markov Decision Process to account for exogenous uncertainties coming from lead time and demand, and integrate it with the scheduling level problem. Shin et al. (2017) use dynamic programming to learn the value function of the Markov Decision Process of a wind farm microgrid. Its value functions penalize the objective functions of the two-stage stochastic programs for daily schedules. Bloch-Mercier (2002) models the deterioration process of a system as continuous-time Markov chain to optimize inspection intervals. Lin et al. (2012) model a simple utility system using Markov chain and carry out RAM (reliability, availability & maintainability) analysis iteratively to decide the optimal reliability design. Chryssaphinou et al. (2011) analyze multi-state reliability systems with discrete-time semi-Markov Chain. Terrazas-Moreno et al. (2010) use Markov chain as an uncertainty modeling tool for the optimal design of production site network considering reliability and flexibility. Kim (2017) presents a reliability model for k-out-of-n systems using a structured continuous-time Markov chain, which is solved with a parallel genetic algorithm.

Given the aforementioned research gaps and knowledge basis, this work extends our recent mixed-integer framework (Ye et al., 2017) and introduces a systematic approach to model the stochastic failure and repair process of the superstructure system as a continuous-time Markov chain. The new framework explicitly accounts for the long term property of each possible reliability scenario. Therefore, it is able to incorporate various kinds of decision making processes. Especially, comparing to Terrazas-Moreno et al. (2010) and Kim (2017), corrective maintenance and condition-based maintenance are incorporated in order to find the overall optimal selection of parallel units.

In section 2, a motivating example is introduced to outline the model scope. Section 3 gives a formal problem statement, including detailed explanation of the modeling assumptions and basic logic, as well as a brief introduction to Markov chain. The mathematical formulation of the model is presented in section 4, while section 5 explains the decomposition and scenario reduction methods. Finally, section 6 demonstrate the model performance with several case studies.

## 2. Motivating example and problem description

Consider an air separation unit (ASU) shown in Figure 1 as a motivating example. Air is fed to a compressor followed by an after-cooler, and then the pre-purifier to remove impurities such as $CO_2$. After that, the air is compressed again by the booster air compressor and cooled by the gas product of nitrogen and liquid product of oxygen. To better utilize the cold utility, the air feed is usually split as follows: about two thirds of the air remains in gas phase, whose temperature is further reduced by a gas turbine before being fed into the high pressure column. The rest of the air is cooled down to be partial-liquid-partial-gas in the heat exchanger, which is then split into two streams and fed into low pressure column and high pressure column separately. It is worth to mention that the liquid $O_2$ comes out from the bottom of the low pressure column, therefore, a pump is needed to bring the stream out, while the liquid $N_2$ product comes out from the high pressure column and does not need to be pressurized.

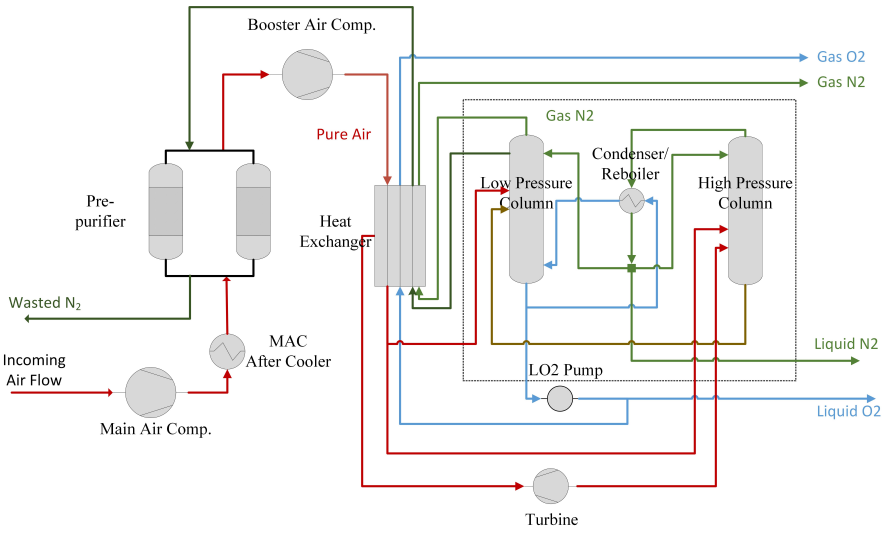



Figure 1: Typical flowsheet of air separation units

The failure of any one of these processing stages can result in the failure of the entire system, which will compromise its ability to meet customer demands. In order to effectively increase the system availability, two strategies are considered.

The first strategy is to install parallel units for the critical stages. In Figure 2, the availability superstructure is formulated as a serial system of sequential stages, where each stage has several potential design alternatives. For example, for the main air compressor, we can install two full-capacity units with one of them as standby, which is more expensive than installing only one unit, but the standby compressor can become active when the primary unit fails and thus avoid losses caused by unavailability.

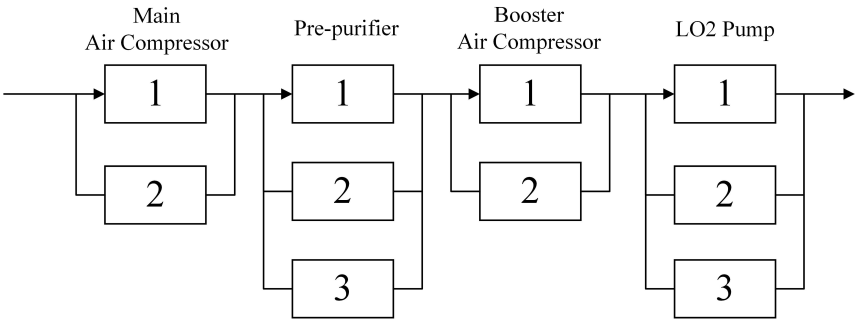



Figure 2: The diagram of ASU reliability design alternatives. Each block represents a parallel unit with certain availability and cost rates

The second strategy is to carry out condition-base maintenance (CBM), which as indicated by its name, is a class of maintenance approach that emphasizes on prioritizing and optimizing maintenance resources with the right strategies for condition monitoring and maintenance. Specifically, the units go through periodic inspections, and follow-up maintenance if the inspection result indicates that the equipment is going to fail shortly. A longer lead time is then allowed for the spare parts to arrive while the equipment is still functional. On the other hand, to repair after the failure actually happens often leads to a longer equipment downtime, or a higher cost to expedite the shipping of spare parts if no spare parts are kept on site. Naturally, there are also costs associated with each inspection or maintenance task.

In summary, there are two levels of trade-offs in this decision making process. The first level is the trade-off between the costs incurred and the revenue gained from the availability increasing strategies. The second level is the allocation of the availability improving budget between the two strategies addressed above and among the processing stages. The ultimate goal is to achieve the optimal overall net present value for the system.

## 3. Problem statement

With the motivating example in mind, we define a general modeling framework for production systems with underlying serial structures for availability evaluation as shown in Figure 3. For each stage $k$, a set of potential parallel units $J_k$ are available for selecting at the design phase. The availability parameters of each unit is known. Our goal is to determine design decisions regarding the number of parallel units to install, as well as operational decisions regarding the length of inspection intervals $t_k^{insp}$. The objective is to optimize the system availability (i.e. probability that the system performs without failures) so as to maximize the profit, i.e. sales revenue minus investment costs and maintenance costs.

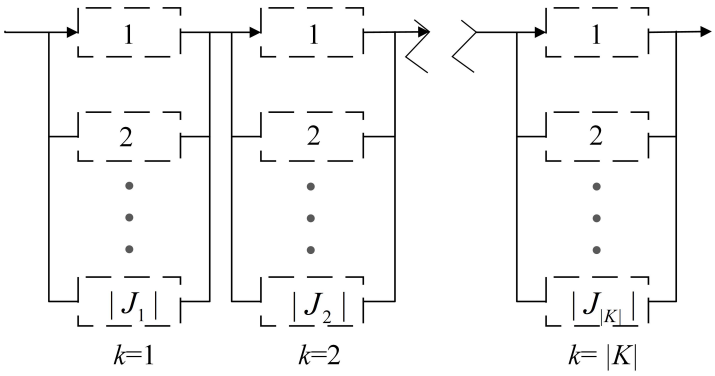



Figure 3: A serial system

Next in section 3.1, the failure and repair processes of single equipment with single failure mode are characterized (for multiple failure modes see Appendix D), based on which the system logic and modeling details will be further discussed in section 3.2. In section 3.3, an example is used to show how Markov chain is applied to the system being studied.

*3.1. The failure and repair processes of single units*

The reliability of a unit is reflected by the random process of its failures, which can be characterized in various ways. One of them is to examine the probability distribution followed by its lifetime, i.e., the time between two consecutive failures. The probability density function is usually referred to as $f(t)$.

A more common characterization is the complementary cumulative distribution function of $f(t)$:

$$R(t) = \int_t^{\infty} f(s)ds = 1 - \int_0^t f(s)ds \tag{1}$$

$R(t)$ can be interpreted as the probability of a unit surviving beyond time point t. A widely accepted assumption applied above is $\int_0^{\infty} f(s)ds = 1$, which means that an equipment will ultimately fail given long enough time.

Based on these two characterizations, one can identify the failure rate, or hazard function of an equipment as:

$$h(t) = \frac{f(t)}{R(t)} \tag{2}$$

The most widely used lifetime distribution in reliability analysis is the Weibull distribution (Weibull et al., 1951), which assumes that the length of the time period between time point 0 and the first failure obeys the following probability density function:

$$f(\lambda, \beta, t) = \lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^{\beta}}, \quad \lambda, \beta, t > 0 \tag{3}$$

$\lambda$ and $\beta$ are denoted as scale and shape parameters, respectively. $\lambda$ indicates the equipment's potential of failure. The greater $\lambda$ is, the faster the equipment tends to fail. $\beta$ indicates the shape of the distribution curve:

- When $\beta = 1$, the Weibull distribution reduces to an exponential distribution:

$$f(t) = \lambda e^{-\lambda t}, \quad R(t) = e^{-\lambda t}$$

where the failure rate does not change with time,

$$h(t) = \lambda$$

- When $\beta \neq 1$,

$$R(t) = e^{-(\lambda t)^{\beta}}$$
$$h(t) = \beta \lambda^{\beta} t^{\beta - 1}$$

If $\beta < 1$, the failure rate decreases with time, but if $\beta > 1$, the failure rate increases with time.

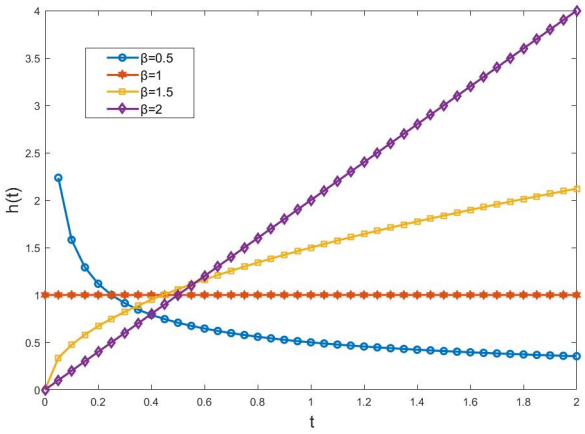



Figure 4: Failure rate against time with $\lambda = 1$ and different $\beta$'s

At the end of section 2, we introduce the strategy that increases production availability by carrying out periodic inspections and follow-up maintenance, with which the units can often be repaired and set back to the beginning of the constant failure period before entering the wear-out period. Therefore, it is reasonable to assume that failure rate does not change with time, and our focus is only on the scale factor $\lambda$:

$$P(lifetime \geq t) = R(t) = e^{-\lambda t} \tag{4}$$

$$f(t) = \lambda e^{-\lambda t} \tag{5}$$

On the basis of an exponential lifetime distribution, the mean time between failure (MTBF) can also be calculated as follows:

$$MTBF = \int_0^{\infty} t f(t) dt = \frac{1}{\lambda} \tag{6}$$

We also know that the failure rate is $\lambda$. For example, if MTBF is 1000 days for a certain piece of equipment, the failure rate will be 0.001.

Once a unit breaks down, repair has to be carried out, which is a more controllable process, though still subject to uncertainties coming from cause detection, spare part shipping, etc. Following from the above discussion of failure processes, the repair process is described by a separate exponential distribution with repair rate $\mu$ as its rate parameter:

$$P(repairtime \geq t) = R(t) = e^{-\mu t} \tag{7}$$

The mean time to repair ($MTTR$) thereby equals to $\frac{1}{\mu}$.

*3.2. Operation rules and decisions identification*

Before presenting the detailed mathematical formulations, we describe in this section the basic logic followed by the system being studied. Each stage $k$ has a set of potential units $j \in J_k$. For unit $j$ in stage $k$, the following information is given:

- Availability parameters without interference, i.e. $\lambda^0_{k,j}$ and $\mu^0_{k,j}$.
- Operating priority level within stage $k$ (indicated by the order of $j$). A unit becomes active if all installed units with higher priority levels have failed.
- Cost data, including installation, inspection, maintenance, and repair.

The parallel units in one stage would come into operation according to their operating priority levels. When a unit fails, a parallel unit with lower operating priority, if it is selected to be installed into the system, might be able to fill in the place to avoid unavailability.

It is assumed that equipment deterioration can be detected by scheduled inspections within a certain period $T^d_k$ before it happens, which is referred to as delay time (Christer, 1999), or PF-interval (Moubray, 1997). As shown in Figure 5, it is called deterioration period in this work. For each stage $k$, inspections are scheduled for active units at a certain time period to be determined, $t^{insp}_k$, called inspection interval. If the inspection indicates that the equipment presents deterioration, a maintenance task will be carried out in time. In that case, catastrophic failures could be prevented, and there will be enough time to order the spare parts, and hence reduce the shipping costs. Therefore, a maintenance before failure causes lower costs and a shorter and more predictable downtime than a repair upon failure. In this paper, we consider the time between failure to be prolonged by condition-based maintenance, and calculate separately the downtime and costs caused by maintenance.

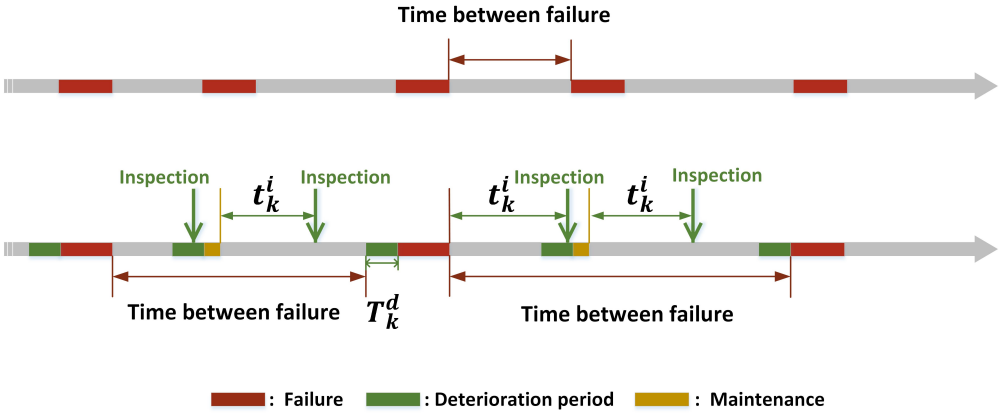
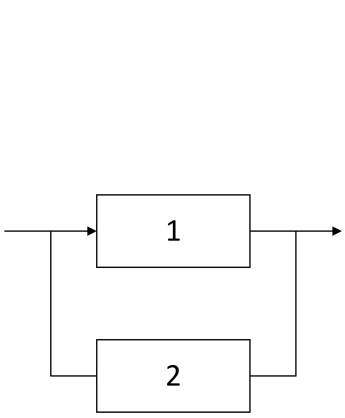



Figure 5: The timeline sketch illustrates how inspection and maintenance tasks affect unit availabilities

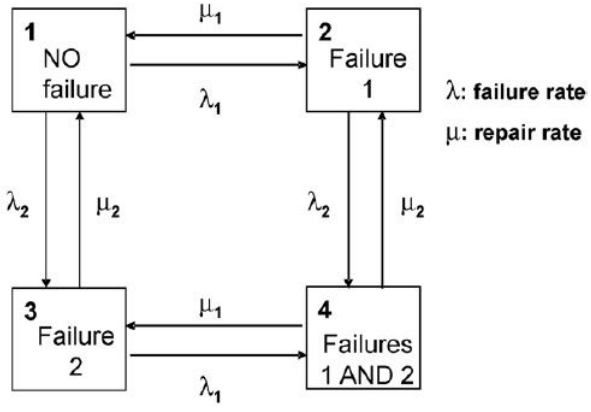
In conclusion, the reliability of stage $k$ depends on the selection of parallel units $y_{k,j}$ and the inspection intervals of each processing stage $t_k^{insp}$.

### *3.3. Example: Construct a Markov chain and solve for the availability*

As the units fail or are repaired, the system being studied transitions randomly among a finite set of states. Moreover, since it is assumed that the time a single unit needs to fail or be repaired follows an exponential distribution, this random state-transitioning process can be represented by a continuous-time Markov chain. The definition and properties of continuous-time Markov chain can be found in Appendix A. In this section, we explain through an example how to represent the Markov chain of a system and solve for its availability. Figure 6(a) shows a system with two independent units 1 and 2. Figure 6(b) shows the state space diagram that include the 4 states, (1)no failure, (2)failure in unit 1, (3)failure in unit 2, and (4)failures in both unit 1 and 2. They constitute the state space $S$ of the Markov chain of this system, which are all possible values of the system state at any time $t$, $X(t)$.




(a) System configuration




(b) State space diagram

Figure 6: A simple 2 unit system

The failure times and repair times of unit 1 and unit 2 follow independent exponential distributions, which is the essential condition for the system to be modeled as a Markov process. $\lambda_1$ and $\lambda_2$ are the failure rates of unit 1 and unit 2 respectively, while $\mu_1$ and $\mu_2$ are the repair rates of unit 1 and unit 2, respectively.

We define matrix $\boldsymbol{P}(t)$ where the element in row $i$, column $j$ $(i, j \in S)$ is denoted as $p_{i,j}(t)$, the probability of the system being in state $j$ at time $t$ (continuous) given that its initial state is $i$, which is described in equation (8). For the case where $i = j$, $p_{i,i}(t)$ is the probability of the system remaining in state $i$ until time point $t$. Therefore, for the example shown in Figure 6, $p_{1,1}(t) = e^{-\lambda_1 t}e^{-\lambda_2 t}$ (see equation (4)). It can be inferred that $p_{i,i}(0) = 1$, $p_{i,j}(0) = 0, \forall j \neq i$, so we have equation (9). In addition, the Markov property requires that the condition in (10) holds for the matrix $\boldsymbol{P}(t)$, where both $s$ and $t$ are time variables.

$$p_{i,j}(t) = Pr\{X(t) = j | X(0) = i\} \tag{8}$$

$$\boldsymbol{P}(0) = \boldsymbol{I}, \quad \lim_{t\to 0}\boldsymbol{P}(t) = \boldsymbol{I} \tag{9}$$

$$p_{i,j}(s+t) = \sum_{k\in S} p_{i,k}(s)p_{k,j}(t) \text{ (Chapman-Kolmogorov equation for continuous Markov chain)} \tag{10}$$

Following from $p_{i,j}(t)$ in (8), $p_{i,j}$ is defined in equation (11),where $T_1$ is the first time point that a state change happens.

$$p_{i,j} = Pr\{X(T_1) = j | X(0) = i\} \tag{11}$$

Still looking at the transitions out of state 1, there is $p_{1,2}/p_{1,3} = \frac{\lambda_1 e^{-\lambda_1 T_1} e^{-\lambda_2 T_1}}{\lambda_2 e^{-\lambda_2 T_1} e^{-\lambda_1 T_1}} = \lambda_1/\lambda_2$ (see equations (4) and (5)), and $p_{1,2} + p_{1,3} = 1$ , thus, we obtain $p_{1,2} = \frac{\lambda_1}{\lambda_1+\lambda_2}$, $p_{1,3} = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

Based on the $\boldsymbol{P}$ matrix, we introduce the transitional rate matrix (transition matrix) $\boldsymbol{Q}$ (the element in row $i$, column $j$ $(i, j \in S)$ of $\boldsymbol{Q}$ is denoted as $q_{i,j}$) defined by equation (12). Again, the special case where $i = j$ is easy to calculate. Following from the expression of $p_{1,1}(t)$, we have $q_{1,1} = -\lambda_1 - \lambda_2$.

$$\boldsymbol{Q} = \boldsymbol{P}'(0) \tag{12}$$

As a matter of fact, $q_{i,j}, i \neq j$ can be calculated with the help of $q_{i,i}$ and $p_{i,j}$ as shown in equation (13) (which is non-trivial to derive, and for proof we refer to Sericola (2013)), which gives us $q_{1,2} = \lambda_1$, $q_{1,3} = \lambda_2$. Similarly, we can obtain the other elements of the transition rate matrix through (14).

$$q_{i,j} = -q_{i,i}p_{i,j}, \quad i \neq j \tag{13}$$

$$\boldsymbol{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} -\lambda_1 - \lambda_2 & \lambda_1 & \lambda_2 & 0 \\ \mu_1 & -\mu_1 - \lambda_2 & 0 & \lambda_2 \\ \mu_2 & 0 & -\mu_2 - \lambda_1 & \lambda_1 \\ 0 & \mu_2 & \mu_1 & -\mu_2 - \mu_1 \end{bmatrix} \end{array} \tag{14}$$

It can be noticed from (14) that $q_{i,j}(i \neq j)$ is equal to the rate parameter from state $i$ to state $j$ marked on the state space diagram of Fig 6, and $q_{i,i}$ is just the opposite number of the sum of all

other elements in row $i$. In fact, the $\boldsymbol{Q}$ matrix is not only very convenient to construct, but also a powerful tool in representing and characterizing the continuous-time Markov process.

The stationary probability of each state $i \in S$, $\pi_i$ is an important measurement of a continuous-time Markov chain, which is also what we pursue throughout this work. For each stationary probability, equation (15) holds. It can be interpreted based on the fact that the realization of a state $j$ is the result of transitions from all states in the space to it. Therefore, the probability of each state $i$ times the transition probability from $i$ to $j$ is the contribution of state $i$ to the probability of a state $j$.

$$\pi_j = \sum_{i \in S} \pi_i p_{i,j}(t), \forall j \in S \tag{15}$$

The matrix form is as shown in equation (16):

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \boldsymbol{P}(t) \tag{16}$$

Now we show how this important equation of $\boldsymbol{\pi}$ based on the probability matrix $\boldsymbol{P}(t)$ can be transformed to a more useful equation of $\boldsymbol{\pi}$ based on the transition rate matrix $\boldsymbol{Q}$. First, we left multiply (12) by $\boldsymbol{\pi}^\top$ to obtain (17)

$$\boldsymbol{\pi}^\top \boldsymbol{Q} = \boldsymbol{\pi}^\top \boldsymbol{P'}(0) = \lim_{t \to 0} \boldsymbol{\pi}^\top \frac{\boldsymbol{P}(t) - \boldsymbol{P}(0)}{t} \tag{17}$$

Substituting (9) into (17) yields (18)

$$\boldsymbol{\pi}^\top \boldsymbol{Q} = \boldsymbol{\pi}^\top \lim_{t \to 0} \frac{\boldsymbol{P}(t) - \boldsymbol{I}}{t} \tag{18}$$

Finally, substituting (16) allows us to obtain equation (19). A qualitative explanation of (19) is that the long-term rate of leaving state $j$, $-\pi_j q_{j,j}$ (notice that $q_{j,j}$ are negative) equals the long-term rate of going into state $j$ from other states, $\sum_{i \neq j} \pi_i q_{i,j}$.

$$\boldsymbol{\pi}^\top \boldsymbol{Q} = \lim_{t \to 0} \frac{\boldsymbol{\pi}^\top - \boldsymbol{\pi}^\top}{t} = 0 \tag{19}$$

In addition, since $\boldsymbol{\pi}$ stands for the probabilities of all possible states, it is required that all of its elements sum to 1. Therefore, after obtaining the $\boldsymbol{Q}$ matrix, we can solve the linear system comprised of equations (20) and (21) for the stationary probability vector $\boldsymbol{\pi}$,

$$\boldsymbol{\pi}^\top \boldsymbol{Q} = \boldsymbol{0} \tag{20}$$

$$\boldsymbol{\pi}^\top \boldsymbol{1} = 1 \tag{21}$$

which is all we need to figure out the system availability. In the example of this section, the stationary probability vector can be expressed analytically with the failure rates and repair rates as follows.

$$\boldsymbol{\pi} = \begin{bmatrix} \frac{\mu_1 \mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ \frac{\lambda_1 \mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ \frac{\lambda_2 \mu_1}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\ \frac{\lambda_1 \lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \end{bmatrix} \tag{22}$$

Thus, the availability of the system is equal to (1 - the probability of state 4), which is $(1 - \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)})$. It can been seen from above that the "$\boldsymbol{Q}$ matrix" is all we need to describe a continuous-time Markov chain. In the following sections, we will analyze how the two availability strategies, which are parallel units and inspections, impact the transition matrix of the system, and thus the system availability.

## 4. Mathematical formulation of the MINLP model

In this section, it is shown how to build the MINLP model for a serial system with $|K|$ stages based on continuous-time Markov Chain. $\lambda_{k,j}$ is the failure rate of unit $j$ in stage $k$, and $\mu_{k,j}$ is the repair rate. We will elaborate on the identification of the transitional rate matrix of a single stage in section 4.2, then in section 4.3 we introduce how to construct the extended CTMC of the entire system using the transitional rate matrix of single stages. In section 4.5, we discuss how the selection of inspection intervals affect equipment failure rates. The relationship between system availability and its profitability will be covered in section 4.6.

### *4.1. Logic constraints*

The logic constraints regarding the design decisions need to be applied. $y_{k,j}$ is the binary variable that indicates whether unit $j$ in stage $k$ is selected. Constraint (23) requires that for each stage at least one unit should be installed.

$$\sum_{j\in J_k} y_{k,j} \geq 1, \quad \forall k \in K \tag{23}$$

### *4.2. CTMC of a single stage*

A single unit can have 3 possible states during the entire time horizon: standby, active, and being repaired. Since the selection of potential units is to be determined, a stage $k$ has a set of potential designs $H_k$. Each design $h \in H_k$ generates a sub state space $T_{k,h}$, which is the set of the combined states of the single units being selected in this design $h$. The union of $T_{k,h}$ of stage $k$ is called the super state space of stage $k$, which is denoted as $S_k$. The word "super" is to distinguish it from an actual irreducible state space where all the states are directly or indirectly connected, while in a super state space, two states from different potential designs have no connecting path.

For example, stage $k$ with three full capacity units shown in Figure 7 has 14 potential states in $S_k$, which has 3 subspace $T_{k,1}$, $T_{k,2}$ and $T_{k,3}$ generated by the three design decisions, respectively. The correspondence between the design decisions and the potential states are shown in Table 1. To keep the main text succinct, it is assumed here that each single unit has one failure scenario called "being repaired". However, the situation of having multiple failure modes can also be captured similarly by duplicating the states with failures shown in Table 1 and replacing "being repaired" with the various failure modes or the combinations of them, which is discussed in Appendix D. The model to be introduced in the following sections can accommodate this extension from single failure mode to multiple failure modes without significant change.

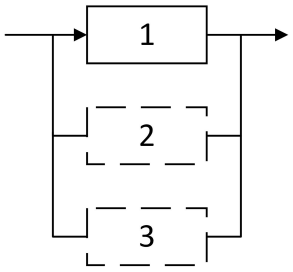
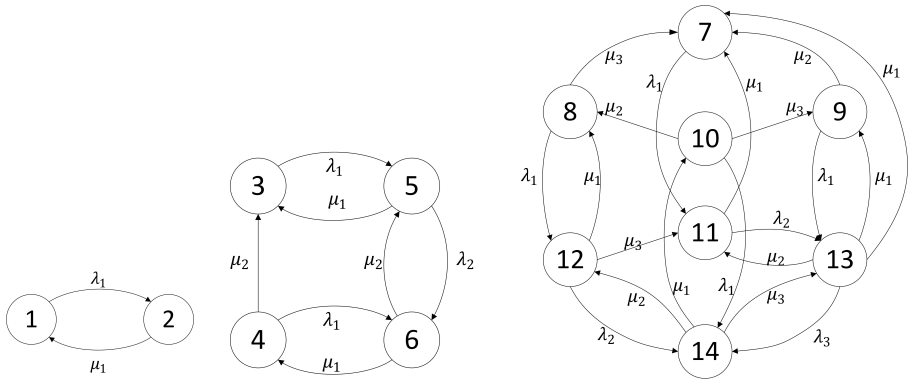



Figure 7: A single stage $k$

Table 1: State enumeration for a stage with identical redundancies

| | Design decisions/subspace | States | unit 1 | unit 2 | unit 3 |
|---|---|---|---|---|---|
| $S_k$ | $T_{k,1}$: unit 1 | state 1 | active | | |
| | | state 2 | being repaired | | |
| | $T_{k,2}$: unit 1 and 2 | state 3 | active | standby | |
| | | state 4 | active | being repaired | |
| | | state 5 | being repaired | active | |
| | | state 6 | being repaired | being repaired | |
| | $T_{k,3}$: unit 1, 2 and 3 | state 7 | active | standby | standby |
| | | state 8 | active | standby | being repaired |
| | | state 9 | active | being repaired | standby |
| | | state 10 | active | being repaired | being repaired |
| | | state 11 | being repaired | active | standby |
| | | state 12 | being repaired | active | being repaired |
| | | state 13 | being repaired | being repaired | active |
| | | state 14 | being repaired | being repaired | being repaired |

As shown in the example in section 3.3, a transitional diagram can be generated for each design decision. The transition rate from a state to its communicating state is equal to the rate parameter of the state changing action. For example, the action that causes the transition from state 1 to state 2 is the failure of unit 1, whose rate parameter is $\lambda_1$. The state transition diagram is as shown in Figure 8, where (a) corresponds to the case where only unit 1 is selected, (b) corresponds to when units 1 and 2 are selected, while (c) is for where all 3 units are selected.




(a) subspace1($T_{k,1}$) (b) subspace2($T_{k,2}$) (c) subspace3($T_{k,3}$)

Figure 8: State transition diagram for the states in Table 1

The state transition diagram of stage $k$ shown in Figure 8 is also reflected in the "super" transition matrix $\boldsymbol{Q}_k$, with the transitions involving failure states in bold. It can be seen that the matrix is block diagonal. Only one of these blocks will become the actual transition matrix of the system, and that is the one selected by the optimization model.

$Q_k =$

| | | $T_{k,1}$ | | $T_{k,2}$ | | | | $T_{k,3}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | **2** | 3 | 4 | 5 | **6** | 7 | 8 | 9 | 10 | 11 | 12 | 13 | **14** |
| $T_{k,1}$ | 1 | $-\lambda_1$ | $\boldsymbol{\lambda_1}$ | | | | | | | | | | | | |
| | **2** | $\boldsymbol{\mu_1}$ | $-\mu_1$ | | | | | | | | | | | | |
| $T_{k,2}$ | 3 | | | $-\lambda_1$ | | $\lambda_1$ | | | | | | | | | |
| | 4 | | | $\mu_2$ | $-\lambda_1-\mu_2$ | | $\boldsymbol{\lambda_1}$ | | | | | | | | |
| | 5 | | | $\mu_1$ | | $-\lambda_2-\mu_1$ | $\boldsymbol{\lambda_2}$ | | | | | | | | |
| | **6** | | | | $\boldsymbol{\mu_1}$ | $\boldsymbol{\mu_2}$ | $-\mu_1-\mu_2$ | | | | | | | | |
| $T_{k,3}$ | 7 | | | | | | | $-\lambda_1$ | | | | $\lambda_1$ | | | |
| | 8 | | | | | | | $\mu_3$ | $-\lambda_1-\mu_3$ | | | | $\lambda_1$ | | |
| | 9 | | | | | | | $\mu_2$ | | $-\lambda_1-\mu_2$ | | | | $\lambda_1$ | |
| | 10 | | | | | | | | $\mu_2$ | $\mu_3$ | $-\lambda_1-\mu_2-\mu_3$ | | | | $\boldsymbol{\lambda_1}$ |
| | 11 | | | | | | | $\mu_1$ | | | | $-\lambda_2-\mu_1$ | | $\lambda_2$ | |
| | 12 | | | | | | | | $\mu_1$ | | | $\mu_3$ | $-\lambda_2-\mu_1-\mu_3$ | | $\boldsymbol{\lambda_2}$ |
| | 13 | | | | | | | $\mu_1$ | | $\mu_1$ | | $\mu_2$ | | $-\lambda_3-2\mu_1-\mu_2$ | $\boldsymbol{\lambda_3}$ |
| | **14** | | | | | | | | | | $\boldsymbol{\mu_1}$ | | $\boldsymbol{\mu_2}$ | $\boldsymbol{\mu_3}$ | $-\mu_1-\mu_2-\mu_3$ |

As discussed and shown above, the existence of subspace $T_{k,h}$ of stage $k$, design $h$, and the block it supports depends on the selection of units $y_{k,j}$. Below we explain how these connections are realized in the model with propositional logic.

Binary variable $z_{k,h}$ (boolean variable $Z_{k,h}$) is defined through the logical proposition (24) to indicate the existence of $T_{k,h}$ based on the values of unit selection binary variable $y_{k,j}$ (boolean variable $Y_{k,j}$ ):

$$Z_{k,h} \Leftrightarrow (\bigwedge_{(j,k,h)\in D} Y_{k,j})(\bigwedge_{(j,k,h)\notin D} \neg Y_{k,j}), \quad \forall k \in K, h \in H_k \tag{24}$$

where $H_k$ is the set of potential designs of stage $k$. $D$ is the set of the index tuples $(j,k,h)$ where design decision $h \in H_k$ includes unit $j \in J_k$ in stage $k$.

For example, if there are 3 potential units in stage 1 ($J_1 = \{1,2,3\}$), then the number of designs is $2^3 - 1 = 7$. The subset of D involving $k$ is $\{(1,k,1), (2,k,1), (1,k,3), (2,k,3), (3,k,4), (1,k,5), (3,k,5), (2,k,6), (3,k,6), (1,k,7), (2,k,7), (3,k,7)\}$.

Table 2: The correspondence of $j$ and $h$ in stage $k$

| | | $h$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $j$ | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| | 3 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Since one and only one of the subspace will be realized, we also have the following logical condition (25):

$$\bigvee_{h\in H_k} Z_{k,h}, \quad \forall k \in K \tag{25}$$

The linear constraints (26) - (28) are translated from the logic condition (24), while (29) is translated from (25). For details of the standard procedure please see Raman and Grossmann (1991).

$$z_{k,h} \leq y_{k,j}, \quad (j,k,h) \in D \tag{26}$$

$$z_{k,h} \leq 1 - y_{k,j}, \quad (j,k,h) \notin D \tag{27}$$

$$z_{k,h} \geq \sum_{(j,k,h)\in D} y_{k,j} + \sum_{(j,k,h)\notin D} (1 - y_{k,j}) - |J_k| + 1, \quad k \in K, h \in H_k \tag{28}$$

$$\sum_{h\in H_k} z_{k,h} = 1, \quad \forall k \in K \tag{29}$$

*4.3. Construction of the extended state space*

After the "super" transitional rate matrices are constructed for single stages $\boldsymbol{Q}_k$ , in this section, the "super" transition matrix of the entire system is calculated based on them. Consider a serial system with $|K|$ stages. Let $n_1$, $n_2$, ..., $n_{|K|}$ be the dimensions of the square transition matrices $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$, ..., $\boldsymbol{Q}_{|K|}$ of the single stages. Let $\boldsymbol{W}$ be the "super" transitional rate matrix of the extended CTMC that describes the stochastic process of the entire system. Following the conventional notations, $I_{(d)}$ are identity matrices of dimension $d$. Following from the results described in Appendix B, we have the formula shown in (30) ,[1]

$$\begin{aligned}
&\boldsymbol{W} = \boldsymbol{I}_{(n_{|K|})} \otimes (\boldsymbol{I}_{(n_{|K|-1})} \otimes \cdots \otimes (\boldsymbol{I}_{(n_3)} \otimes (\boldsymbol{I}_{(n_2)} \otimes \boldsymbol{Q}_1 + \boldsymbol{Q}_2 \otimes \boldsymbol{I}_{(n_1)}) + \boldsymbol{Q}_3 \otimes \boldsymbol{I}_{(n_1 n_2)}) + \cdots + \boldsymbol{Q}_{|K|-1} \otimes \boldsymbol{I}_{(n_1 n_2 \ldots n_{|K|-2})}) \\
&+ \boldsymbol{Q}_{|K|} \otimes \boldsymbol{I}_{(n_1 n_2 \ldots n_{|K|-1})}) \\
&= \boldsymbol{I}_{(n_{|K|} n_{|K|-1} \ldots n_2)} \otimes \boldsymbol{Q}_1 + \boldsymbol{I}_{(n_{|K|} n_{|K|-1} \ldots n_3)} \otimes \boldsymbol{Q}_2 \otimes \boldsymbol{I}_{(n_1)} + \boldsymbol{I}_{(n_{|K|} n_{|K|-1} \ldots n_4)} \otimes \boldsymbol{Q}_3 \otimes \boldsymbol{I}_{(n_2 n_1)} + \cdots + \\
&\boldsymbol{I}_{(n_{|K|})} \otimes \boldsymbol{Q}_{|K|-1} \otimes \boldsymbol{I}_{(n_{|K|-2} n_{|K|-3} \ldots n_1)} + \boldsymbol{Q}_{|K|} \otimes \boldsymbol{I}_{(n_{|K|-1} n_{|K|-2} \ldots n_1)}
\end{aligned} \tag{30}$$

With $S_k$ representing the state space of stage $k \in K$, let $\bar{S}$ be the state space of the extended CTMC. According to the properties of the Kronecker product, we have,

$$|\bar{S}| = \prod_{k \in K} |S_k| \tag{31}$$

A system state $\bar{s} \in \bar{S}$ is the combination of $|K|$ stage states $s \in S_k$ from every processing stage $k \in K$. For example, consider a system with two stages shown in Figure 9, where unit 1 is selected for stage 1 and units 1&2 for stage 2 (units of solid lined boxes in shades are selected).

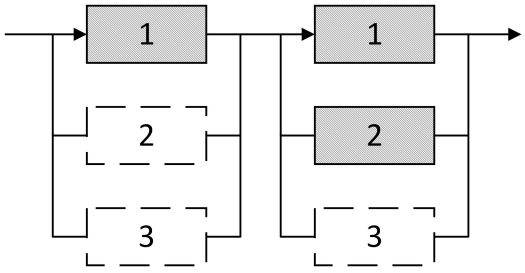



Figure 9: A single stage $k$

The state space of stage 1, $S_1 = \{1, 2\}$, and the state space of stage 2, $S_2 = \{3, 4, 5, 6\}$ (refer to Table 1). The cardinality of the extended state space, $|\bar{S}| = 8$. State $\bar{s} = 29$ for example, means having state 1 in $S_1$, and state 3 in $S_2$. The corresponding relationships for the entire set are displayed in Figure 10.

[1]"$\otimes$" is an operation on two matrices of arbitrary dimensions. $\boldsymbol{A} \otimes \boldsymbol{B}$ is called the Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. If $\boldsymbol{A}$ is of dimension $a_1 \times a_2$, $\boldsymbol{B}$ is of dimension $b_1 \times b_2$, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is of dimension $a_1 b_1 \times b_1 b_2$.

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} A_{11}B & A_{12}B & \cdots & A_{1a_2}B \\ \vdots & \vdots & & \vdots \\ A_{a_1 1}B & A_{a_1 2}B & \cdots & A_{a_1 a_2}B \end{bmatrix}$$

Kronecker product satisfies the associative law of addition:

$$\boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{A} \otimes \boldsymbol{C} = \boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C})$$

| Super state space: $\bar{S}$ | | State space of stage 2: $S_2$ | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| State space of stage 2: $S_1$ | 1 | 1 | 15 | **29** | **43** | **57** | **71** | 85 | 99 | 113 | 127 | 141 | 155 | 169 | 183 |
| | 2 | 2 | 16 | **30** | **44** | **58** | **72** | 86 | 100 | 114 | 128 | 142 | 156 | 170 | 184 |
| | 3 | 3 | 17 | 31 | 45 | 59 | 73 | 87 | 101 | 115 | 129 | 143 | 157 | 171 | 185 |
| | 4 | 4 | 18 | 32 | 46 | 60 | 74 | 88 | 102 | 116 | 130 | 144 | 158 | 172 | 186 |
| | 5 | 5 | 19 | 33 | 47 | 61 | 75 | 89 | 103 | 117 | 131 | 145 | 159 | 173 | 187 |
| | 6 | 6 | 20 | 34 | 48 | 62 | 76 | 90 | 104 | 118 | 132 | 146 | 160 | 174 | 188 |
| | 7 | 7 | 21 | 35 | 49 | 63 | 77 | 91 | 105 | 119 | 133 | 147 | 161 | 175 | 189 |
| | 8 | 8 | 22 | 36 | 50 | 64 | 78 | 92 | 106 | 120 | 134 | 148 | 162 | 176 | 190 |
| | 9 | 9 | 23 | 37 | 51 | 65 | 79 | 93 | 107 | 121 | 135 | 149 | 163 | 177 | 191 |
| | 10 | 10 | 24 | 38 | 52 | 66 | 80 | 94 | 108 | 122 | 136 | 150 | 164 | 178 | 192 |
| | 11 | 11 | 25 | 39 | 53 | 67 | 81 | 95 | 109 | 123 | 137 | 151 | 165 | 179 | 193 |
| | 12 | 12 | 26 | 40 | 54 | 68 | 82 | 96 | 110 | 124 | 138 | 152 | 166 | 180 | 194 |
| | 13 | 13 | 27 | 41 | 55 | 69 | 83 | 97 | 111 | 125 | 139 | 153 | 167 | 181 | 195 |
| | 14 | 14 | 28 | 42 | 56 | 70 | 84 | 98 | 112 | 126 | 140 | 154 | 168 | 182 | 196 |

Figure 10: Example on the correspondence between stage state $s \in S_k$ and system state $\bar{s} \in \bar{S}$

As indicated in Figure 10, the index of $\bar{s}$ in $\bar{S}$ are arranged in the order of the corresponding indices $s$ in $S_k$, with larger $k$ as prior dimensions. $SC$ is the set of the index tuples where $s$ in $S_k$ corresponds to $\bar{s}$ in $\bar{S}$, which can be calculated by (32).

$$SC = \{(k, s, \bar{s}) | s = \lceil \frac{mod(\bar{s} - 1, \prod_{l \in K, l \leq k} |S_l|) + 1}{\prod_{l \in K, l \leq k-1} |S_l|} \rceil, \quad k \in K, s \in S_k, \bar{s} \in \bar{S}\} \tag{32}$$

Particularly, the set of failed states in stage $k$ is denoted as $S_k^f$, and similarly, the set of system states that fails is denoted as $\bar{S^f}$.

$$\bar{S^f} = \{\bar{s} \in \bar{S} | \forall k \in K, \quad \exists s \in S_k^f, \text{ s.t. } (k, s, \bar{s}) \in SC\} \tag{33}$$

Continuing with the example shown in Figure 9, the transition matrix of the two stages are $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ as shown below. The transition matrix of the extended CTMC is $\boldsymbol{W}$.

$$\boldsymbol{Q}_1 = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} -\lambda_{1,1} & \lambda_{1,1} \\ \mu_{1,1} & -\mu_{1,1} \end{bmatrix} \end{array}, \quad \boldsymbol{Q}_2 = \begin{array}{c} \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} -\lambda_{2,1} & & \lambda_{2,1} & \\ \mu_{2,2} & -\lambda_{2,1} - \mu_{2,2} & & \lambda_{2,1} \\ \mu_{2,1} & & -\lambda_{2,2} - \mu_{2,1} & \lambda_{2,2} \\ & \mu_{2,1} & \mu_{2,2} & -\mu_{2,1} - \mu_{2,2} \end{bmatrix} \end{array}$$

$$\boldsymbol{W} = \boldsymbol{I}_{(4)} \otimes \boldsymbol{Q}_1 + \boldsymbol{Q}_2 \otimes \boldsymbol{I}_{(2)} =$$

$$\begin{array}{c} \\ (1,3) \\ (2,3) \\ (1,4) \\ (2,4) \\ (1,5) \\ (2,5) \\ (1,6) \\ (2,6) \end{array} \begin{array}{c} \begin{array}{cccccccc} (1,3) & (2,3) & (1,4) & (2,4) & (1,5) & (2,5) & (1,6) & (2,6) \end{array} \\ \begin{bmatrix} -\lambda_{1,1} - \lambda_{2,1} & \lambda_{1,1} & & & \lambda_{2,1} & & & \\ \mu_{1,1} & -\lambda_{2,1} - \mu_{1,1} & & & & \lambda_{2,1} & & \\ \mu_{2,2} & & -\lambda_{1,1} - \lambda_{2,1} - \mu_{2,2} & \lambda_{1,1} & & & \lambda_{2,1} & \\ & \mu_{2,2} & \mu_{1,1} & -\lambda_{2,1} - \mu_{1,1} - \mu_{2,2} & & & & \lambda_{2,1} \\ \mu_{2,1} & & & & -\lambda_{1,1} - \lambda_{2,2} - \mu_{2,1} & \lambda_{1,1} & \lambda_{2,2} & \\ & \mu_{2,1} & & & \mu_{1,1} & -\lambda_{2,2} - \mu_{1,1} - \mu_{2,1} & & \lambda_{2,2} \\ & & \mu_{2,1} & & \mu_{2,2} & & -\lambda_{1,1} - \mu_{2,1} - \mu_{2,2} & \lambda_{1,1} \\ & & & \mu_{2,1} & & \mu_{2,2} & \mu_{1,1} & -\mu_{1,1} - \mu_{2,1} - \mu_{2,2} \end{bmatrix} \end{array}$$

Directly establishing the connection between stage states and system states in the model can result in a very large number of equations. Therefore, we circumvent this problem by only connecting the stage sub state spaces and system sub state spaces (see Figure 11)

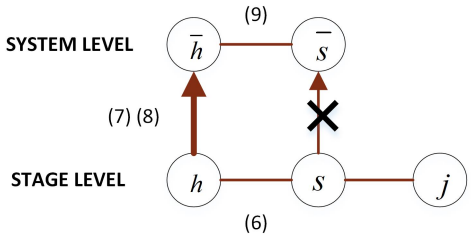



Figure 11: The relationship between the stage-wise indices and the system-wise indices

As shown in Table 1, stage $k$ has several mutually exclusive subspaces denoted by $T_{k,h}$:$\{1,2\}$, $\{3,4,5,6\}$, $\{7,8,9,10,11,12,13,14\}$, whose existence is determined by the unit selection $y_{k,j}$ of stage $k$. Accordingly, we define mutually exclusive subspaces $\bar{T}_{\bar{h}}$ of the system, which are each generated by a different system design $\bar{h}$.

As mentioned above, each system state $\bar{s} \in \bar{S}$ corresponds to a state $s \in S_k$ from each stage $k$. Similarly, each system design $\bar{h} \in \bar{H}$ has a corresponding design $h \in H_k$ from each stage $k$. Moreover, the system subspace $\bar{T}_{\bar{h}}$ generated by design $\bar{h}$ contains those system states $\bar{s}$ such that, for each stage $k$, the corresponding state $s$ of $\bar{s}$ belongs to the subspace $T_{k,h}$ generated by $\bar{h}$'s corresponding stage design $h$.

Similarly to $SC$ in equation (32), we can calculate the set $HC$ through (34):

$$HC = \{(k,h,\bar{h}) | h = \lceil \frac{mod(\bar{h}-1, \prod_{l\in K, l\leq k} |H_l|) + 1}{\prod_{l\in K, l\leq k-1} |H_l|} \rceil, \quad k \in K, h \in H_k, \bar{h} \in \bar{H}\} \tag{34}$$

The mathematical definition of the system state subspace $\bar{T}_{\bar{h}}$ is then as defined in equation (35):

$$\bar{T}_{\bar{h}} = \{\bar{s} \in \bar{S} | \forall k \in K, \quad \exists h \in H_k, s \in T_{k,h}, \text{ s.t. } (k,h,\bar{h}) \in HC, (k,s,\bar{s}) \in SC\} \tag{35}$$

Now we consider again the system shown in Figure 9, but without the design decisions specified. Thus, $H_1 = \{1,2,3\}$, $H_2 = \{1,2,3\}$, $\bar{H} = \{1,2,3,4,5,6,7,8,9\}$

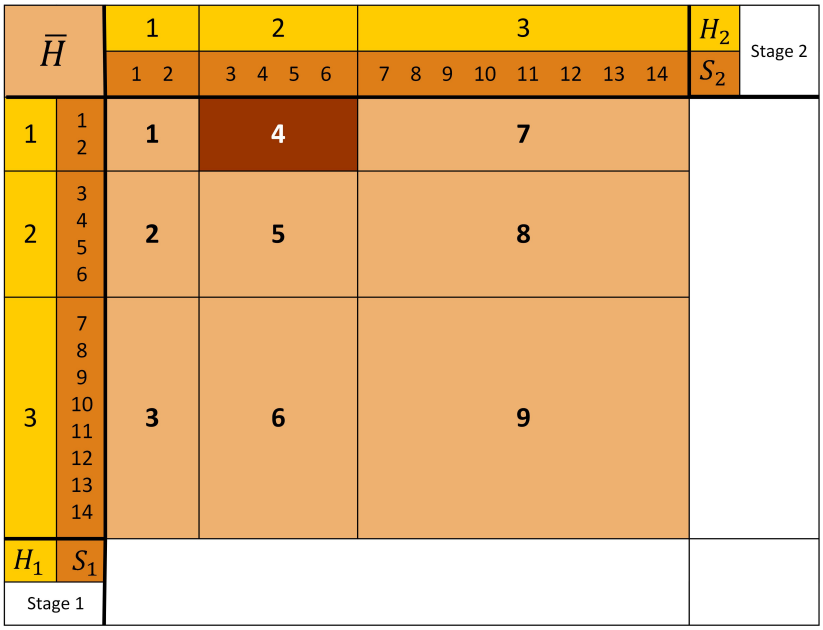



Figure 12: Example on the correspondence between $h \in H_k$ and $\bar{h} \in \bar{T}$

The system states are grouped into mutually exclusive blocks $\bar{T}_{\bar{h}}$ whose existence are subject to the subspaces $T_{k,h}$ of single stages $k$. Only one of the nine large blocks in the middle will be the actual state space of the two stage system we have been discussing. For example, in this case where the design is "unit 1 for stage 1 and unit 1&2 for stage 2", then block 4 is the actual state space.

The existence of $T_{k,h}$, which is indicated by the binary variable $z_{k,h}$, determines that of $\bar{T}_{\bar{h}}$. The logical condition between $\bar{z}_{\bar{h}}$ and $z_{k,h}$ is expressed as follows,

$$\bar{Z}_{\bar{h}} \Leftrightarrow \bigwedge_{(k,h,\bar{h}) \in HC} Z_{k,h} \tag{36}$$

(37) and (38) are the linear constraints reformulated from the logical condition (36). For details of the standard translation procedure please see Raman and Grossmann (1991).

$$\bar{z}_{\bar{h}} \leq z_{k,h}, \quad \forall (k,h,\bar{h}) \in HC \tag{37}$$

$$\bar{z}_{\bar{h}} \geq \sum_{k \in K} z_{k,h} - |K| + 1, \quad \forall (k,h,\bar{h}) \notin HC \tag{38}$$

Equation (39) requires that subspace $\bar{T}_{\bar{h}}$ exists simultaneously with all its elements.

$$\overline{zz}_{\bar{s}} = \bar{z}_{\bar{h}}, \quad \forall \bar{s} \in \bar{T}_{\bar{h}} \tag{39}$$

*4.4. Solve for stationary probability vector*

As described in section 3.3, the "super" transition rate matrix of the entire system, $\boldsymbol{W}$, is used to calculate the stationary probabilities $\boldsymbol{\pi}$ through the following linear system (40), which is adapted

from equations (20) and (21) with the generic notation for transition matrix, $\boldsymbol{Q}$ replaced with $\boldsymbol{W}$:

$$\boldsymbol{\pi}^\top \boldsymbol{W} = 0 \tag{40}$$

$$\boldsymbol{\pi}^\top \mathbf{1} = 1 \tag{41}$$

Disjunction (42) requires that if a system state $\bar{s}$ does not exist, its stationary probability $\pi_{\bar{s}}$ should be zero, otherwise, it is less than or equal to 1:

$$\begin{bmatrix} \overline{ZZ}_{\bar{s}} \\ 0 \leq \pi_{\bar{s}} \leq 1 \end{bmatrix} \vee \begin{bmatrix} \neg\overline{ZZ}_{\bar{s}} \\ \pi_{\bar{s}} = 0 \end{bmatrix} \tag{42}$$

(42) is translated into the inequality (43).

$$\pi_{\bar{s}} \leq \overline{zz}_{\bar{s}}, \quad \bar{s} \in \bar{S} \tag{43}$$

Having the above constraint is equivalent to eliminating the non-existing rows in $\boldsymbol{W}$. In addition to that, the columns in $\boldsymbol{W}$ corresponding to the non-existing states need to be eliminated, which means that the corresponding equations in the linear system $\boldsymbol{\pi}^\top \boldsymbol{W} = 0$ need to be relaxed. Let $\bar{s}$ be row index and $\bar{r}$ be column index of $\boldsymbol{W}$, we have disjunction (44),

$$\begin{bmatrix} \overline{ZZ}_{\bar{r}} \\ \sum_{\bar{s}} \pi_{\bar{s}} W(\bar{s}, \bar{r}) = 0 \end{bmatrix} \vee \begin{bmatrix} \neg\overline{ZZ}_{\bar{r}} \\ \sum_{\bar{s}} \pi_{\bar{s}} W(\bar{s}, \bar{r}) < \infty \end{bmatrix} \tag{44}$$

which is translated into constraints (45) - (46) through the big-M reformulation (Grossmann and Trespalacios, 2013),

$$\sum_{\bar{s}} \pi_{\bar{s}} W(\bar{s}, \bar{r}) \leq M(1 - \overline{zz}_{\bar{r}}), \quad \forall \bar{r} \in \bar{R} \tag{45}$$

$$\sum_{\bar{s}} \pi_{\bar{s}} W(\bar{s}, \bar{r}) \geq M(\overline{zz}_{\bar{r}} - 1), \quad \forall \bar{r} \in \bar{R} \tag{46}$$

Figure 13 sketches what happens to the non-existing rows and columns in $\boldsymbol{W}$.

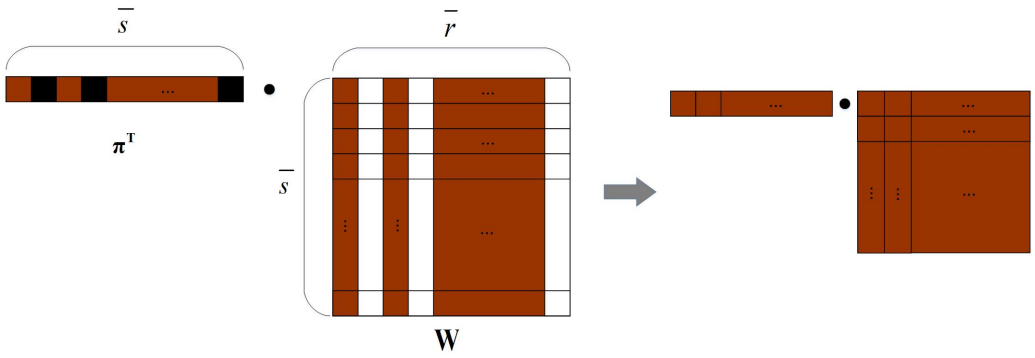



Figure 13: Row and column eliminations. Equation (43) removes the black blocks, while Equations (45) - (46) remove the white columns

Finally, the availability of the system is one minus the sum of the stationary probability of all failed states:

$$A = 1 - \sum_{\bar{s} \in \bar{S}^f} \pi_{\bar{s}} \tag{47}$$

*4.5. Inspections and maintenance*

In the previous sections, a modeling framework has been presented based on Markov process, focusing on the impact of system configuration on production availability, where failure rates and repair rates serve as key reliability parameters. However, the reliability parameters can be varied by operational factors such as maintenance activities. Especially, the idea of condition-based maintenance is to carry out maintenance tasks according to equipment conditions, which depends on real-time condition monitoring and periodical inspections. If a unit is maintained more frequently, its failure rates ($\lambda$) could decrease. In fact, in practical plant operations, carrying out maintenance is an effective strategy for improving the reliability of the system and thus, profitability. However, they also add up to operational costs. Therefore, in order to determine the overall optimum, it is essential to assess the impact of maintenance policy on the failure rates.

Recall the discussion in section 3.2, equipment deterioration can be detected by scheduled inspections within a certain time period $T_k^d$ before a failure occurs. For each stage $k$, inspections are scheduled for active equipment at a certain time interval $t_k^{insp}$ to be determined, called inspection intervals, which is recounted after each inspection, maintenance or repair. If the inspection result indicates that the equipment will fail shortly, then a maintenance task will be carried out in time.

The time needed for a repair process is subject to the shipping of spare parts. However, if the failure is predicted and a maintenance task is carried out before it happens, there will be enough time to order the spare parts and hence, reduce the costs.

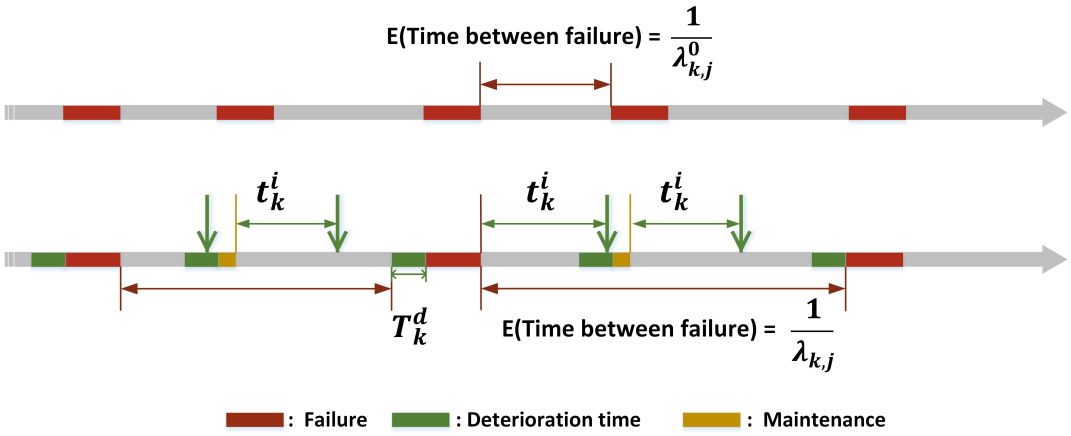



Figure 14: The timeline sketch illustrates how inspections and maintenance affect the availability of unit $j$ in stage $k$

Figure 14 shows the impact of inspection-maintenance on the equivalent failure rates. With periodical inspection, some of the potential failures can be detected during the deterioration time (green bars), so that maintenance can be scheduled accordingly and avoid the failure. Therefore, the time between failure can be prolonged from $\frac{1}{\lambda^0_{k,j}}$ to $\frac{1}{\lambda_{k,j}}$, and the failure rate is reduced from $\lambda^0_{k,j}$ to $\lambda_{k,j}$. The more frequent inspections are carried out, the more the failure rate is expected to be reduced.

Let the "time to failure" of a unit after last inspection, maintenance or repair be random

variable $t^f$. Then, the probability of an inspection being successful is given by equation (48),

$$P(t_k^{insp} \leq t^f \leq t_k^{insp}+T_k^d) = \int_{t_k^{insp}}^{t_k^{insp}+T_k^d} \lambda_{k,j}^0 e^{-\lambda_{k,j}^0 t} dt = e^{-\lambda_{k,j}^0 t_k^{insp}} - e^{-\lambda_{k,j}^0 (t_k^{insp}+T_k^d)}, \quad \forall k \in K, j \in J_k \tag{48}$$

where $\lambda_{k,j}^0$ is the original failure rate of equipment. The expected number of successful inspections during the entire time horizon $T$ is then given by (49),

$$(e^{-\lambda_{k,j}^0 t_k^{insp}} - e^{-\lambda_{k,j}^0 (t_k^{insp}+T_k^d)}) \frac{T}{t_k^{insp}} \tag{49}$$

where $\lambda_{k,j}^0$ is the original failure rate of equipment and $\lambda_{k,j}$ is the equivalent failure rate under the impact of inspections and maintenance, which are used in the transitional rate matrices mentioned in previous sections.

If an inspection result indicates that a failure is going to happen, then the unit will undergo a maintenance to avoid the unplanned failure and repair later on. Thus, as shown in equation (50) the original expected number of repairs, $\lambda_{k,j}^0 T$, minus the expected number of successful inspections gives the expected number of repairs that actually happen, $\lambda_{k,j} T$:

$$\lambda_{k,j}^0 T - (e^{-\lambda_{k,j}^0 t_k^{insp}} - e^{-\lambda_{k,j}^0 (t_k^{insp}+T_k^d)}) \frac{T}{t_k^{insp}} = \lambda_{k,j} T, \quad \forall k \in K, j \in J_k \tag{50}$$

Following from the above equation, the equivalent failure rate $\lambda_{k,j}$ of unit $j$ in stage $k$ is represented as in equation (51). It can be seen that the shorter the inspection interval $t_k^{insp}$ is, the more the equivalent failure rate $\lambda_{k,j}$ is reduced compared to the original failure rate $\lambda_{k,j}^0$.

$$\lambda_{k,j}^0 - \lambda_{k,j} = (e^{-\lambda_{k,j}^0 t_k^{insp}} - e^{-\lambda_{k,j}^0 (t_k^{insp}+T_k^d)}) / t_k^{insp}, \quad \forall k \in K, j \in J_k \tag{51}$$

Considering the executability of the inspection plan as well as to simplify the model formulation, the range of possible inspection intervals $t_k^{insp}$ is discretized into a finite set of choices $T_l^{insp}, l \in L$. The selection of inspection intervals for each stage $k$ is represented with binary variables $x_{k,l}$, where $x_{k,l} = 1$ when time length $T_l^{insp}$ is selected for stage $k$. Equation (52) requires that exactly one choice is selected for each stage. Equation (53) shows that the value of $t_k^{insp}$ is expresses in terms of the values of $T_l^{insp}$ depending on which one is selected.

$$\sum_{l \in L} x_{k,l} = 1, \forall k \in K \tag{52}$$

$$t_k^{insp} = \sum_{l \in L} x_{k,l} T_l^{insp}, \forall k \in K \tag{53}$$

With (52) and (53), equation (51) can be rewritten as (54), which is linear.

$$\lambda_{k,j}^0 - \lambda_{k,j} = \sum_{l \in L} x_{k,l} (e^{-\lambda_{k,j}^0 T_l^{insp}} - e^{-\lambda_{k,j}^0 (T_l^{insp}+T_k^d)}) / T_l^{insp}, \quad \forall k \in K, j \in J_k \tag{54}$$

Each stage $k$ has its own inspection method that suits the equipment, and hence a corresponding cost $c\_insp_k$. Equation (55) The inspection cost for stage $k$ is proportional to its inspection frequency and cost rate.

$$inspCost = \sum_{k \in K} c\_insp_k \sum_{l \in L} x_{k,l} \frac{T}{T_l^{insp}} \tag{55}$$

The repair cost is calculated according to the failure states. The repair cost in each state $\bar{s} \in \bar{S}$ is equal to the frequency of $\bar{s}$, $-W(\bar{s}, \bar{s})\pi_{\bar{s}}$, where $W(\bar{s}, \bar{s})$ are the diagonal elements of the transition matrix, which are negative (see equation (14)), times the summation of the repair costs of all the units that are failed in state $\bar{s}$. Since $W(\bar{s}, \bar{s})$ is subject to the equivalent failure rates of single units, the state frequency, $-W(\bar{s}, \bar{s})\pi_{\bar{s}}$ is a bilinear term, introducing non-linearity in the model.

$$repaCost = -T \sum_{\bar{s} \in \bar{S}} W(\bar{s}, \bar{s})\pi_{\bar{s}} \sum_{k,j \in K J_{\bar{s}}^f} c\_repa_k \tag{56}$$

The number of times for follow-up maintenance to take place in a single unit relative to its number of repairs is calculated by the relative difference between the equivalent failure rate and the original failure rate. In equation (57) we let the maximum relative number of maintenance times among all selected units(indicated by binary variable $y_{k,j}$) be the relative number of maintenance times of the stage.

$$mainRatio_k \geq y_{k,j}(\lambda_{k,j}^0 - \lambda_{k,j})/\lambda_{k,j}, \quad \forall j \in J_k \tag{57}$$

(58) follows the same logic as in (56) to calculate costs according to failure states. Here, $c\_repa_k$ is replaced by $c\_main_k$ times $mainRatio_k$, which is the number of follow-up maintenance relative to the number of repairs.

$$mainCost = -T \sum_{\bar{s} \in \bar{S}} W(\bar{s}, \bar{s})\pi_{\bar{s}} \sum_{k,j \in K J_{\bar{s}}^f} mainRatio_k c\_main_k \tag{58}$$

In addition to the costs, maintenance also causes downtime, which will result in the decrease of availability. Similar but slightly different from equation (58), equation (59) calculates the downtime caused by maintenance in terms of the failure states $\bar{s} \in \bar{S}^f$ and those stages that fail in $\bar{s}$. The net system availability $A^{net}$ is calculated as $A$ minus the ratio of downtime caused by maintenance to the entire time horizon $T$.

$$mainTime = -T \sum_{\bar{s} \in \bar{S}^f} W(\bar{s}, \bar{s})\pi_{\bar{s}} \sum_{k \in K_{\bar{s}}^f} mainRatio_k T\_main_k \tag{59}$$

$$A^{net} = A - \frac{mainTime}{T} \tag{60}$$

*4.6. Objective function: income and expenses*

As shown in equation (61), the objective to be maximized is the Net Present Value $NPV$, which is the present value of net cash flow minus the investment costs. The yearly net cash flow is equal to revenue ($RV$) minus penalty ($PN$), plus bonus ($BN$), minus all the operational costs, and divided by number of years, $\tau$, in the entire time horizon $T$. It is discounted by $\frac{1}{(1+r)^i}$, where

$r$ is the rate of return(RoR) of cash flow. Note that the sum of the discount factors, $\sum_{i=1}^{\tau} \frac{1}{(1+r)^i}$, reduces to the general form, $\frac{1-(1+r)^{-\tau}}{r}$.

$$max \quad NPV = \frac{1}{\tau}(RV - PN + BN - repaCost - inspCost - mainCost) \cdot [\frac{1-(1+r)^{-\tau}}{r}] - instCost \tag{61}$$

where $instCost$ is the investment cost for installing the units depending on binary variables $y_{k,j}$,

$$instCost = \sum_{k} \sum_{j} y_{k,j} c\_inst_{k,j} \tag{62}$$

and the other costs, i.e., $repaCost$, $inspCost$ and $mainCost$ are already described in section 4.5.

The revenue, $RV$, penalty, $PN$ and bonus, $BN$ are impacted by the system availability as discussed in Ye et al. (2017). For convenience of the readers, the detailed equations and explanations can be found in Appendix C.

In conclusion, the MINLP model optimizes (61) subject to equations or inequalities (23), (26)-(29), (37)-(39), (43), (45)-(47), (52)-(60), (62), (C.1)-(C.2) and (C.8)- (C.12), with nonlinearities in equations (45)-(46), and (56) -(59)

## 5. Model decomposition and scenario reduction

An advantage of the model presented above is that it explicitly accounts for every possible state in the system. However, it can also become a drawback when it comes to computational efficiency, especially with the failure rates treated as variables of inspection decisions. For example, the number of variables and equations are already in the order of $10^{12}$ when a system of 4 stages each with 3 distinct parallel units is considered. Therefore, we propose the following decomposition and reduction methods to control the model complexity.

### *5.1. Model decomposition*

As shown in Figure 15, the model can be decomposed by the two types of independent decisions, design decisions and maintenance decisions. In the MINLP with fixed design decisions, the dimension of the system transition matrix is greatly reduced, while in the MILP with fixed inspection intervals, the entries of system transition matrix become parameters, therefore the non-linearities are removed. Thus, solving any of the subproblems requires much less effort than directly solving the original MINLP.

In the initialization stage, the most and the least frequent inspection plans are determined by solving the reduced MINLP with the design decisions fixed to the fewest number of units and the full design, respectively. The most frequent inspection plan is used to calculate a group of failure rates that enter the matrix $W$ in equations (45), (46), (56), (58) and (59)). The least frequent inspection plan is used in maintenance cost calculation ($x_{k,l}$, $t_k^{insp}$, $\lambda_{k,j}$ and $mainRatio_k$ in equations (52) - (55), and (57) - (59)). With these parameters fixed, the model becomes an MILP. This MILP model not only gives an optimistic estimation (upper bound) of the net present value, but also balances between reliability and cost consideration the influence of the inspection-maintenance strategy on the unit selection.

In each iteration, the MILP is solved to generate a design plan and provide an upper bound. With the design decision fixed at this point, a much smaller MINLP formulation is obtained and solved to obtain a feasible solution and a lower bound of the objective function. Unless the gap is within a specified tolerance $\epsilon$, the iteration continues by adding integer cuts of the previously selected designs to the upper bound generating MILP.

The integer cuts are applied on the unit selection level. Assume $m$ iterations have been completed, and all the previously selected designs are represented by $\hat{y}_{m,k,j}$. $\hat{y}_{m,k,j} = 1$ means that unit $j$ of stage $k$ is selected by the MILP solved at iteration $m$. The cuts that are included in the MILP at iteration $m+1$ is

$$\sum_{k\in K}\sum_{j\in J_k,\hat{y}_{i,k,j}=0} y_{k,j} + \sum_{k\in K}\sum_{j\in J_k,\hat{y}_{i,k,j}=1} (1-y_{k,j}) \geq 1, \quad i \leq m \tag{63}$$

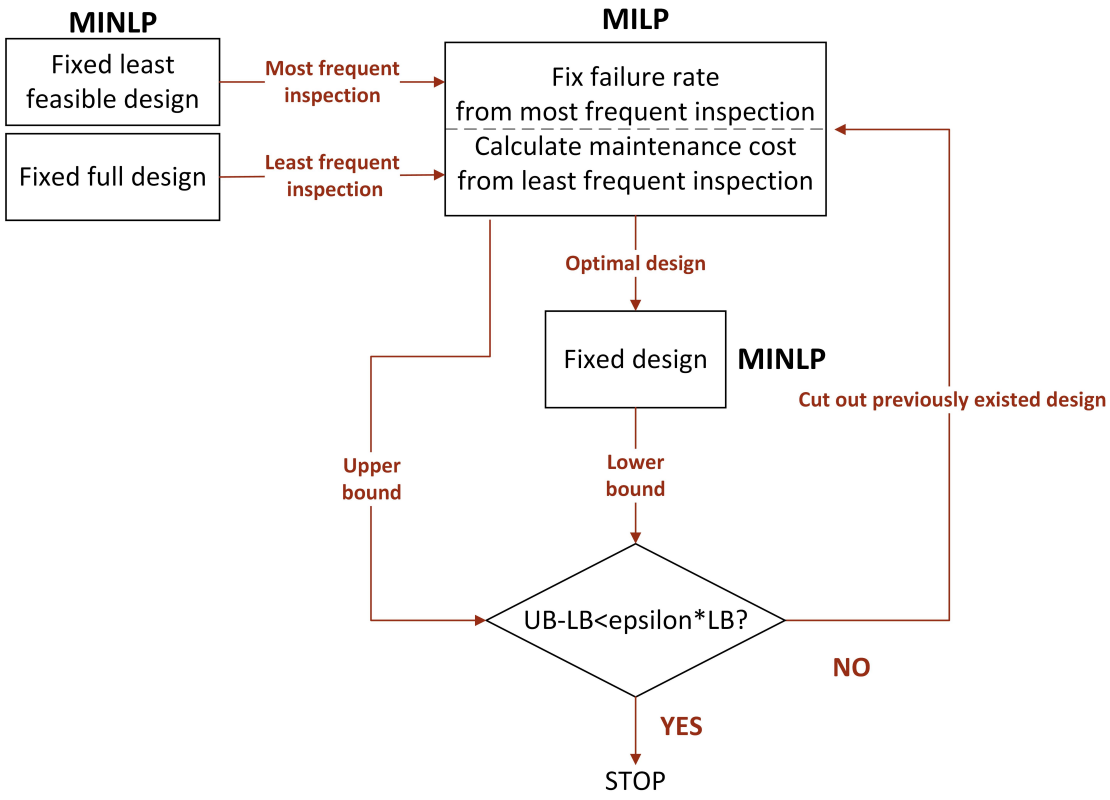



Figure 15: Decomposition framework

*5.2. State space reduction*

By excluding states with very low probabilities, the size of the transition rate matrix of the system can be further reduced without compromising the accuracy of the solution of the model. To be specific, recall that in section 4.2, we show in an example how the state spaces are constructed, which covers all possible states with positive probabilities when certain design is selected. In particular, there are 8 possible states when all three potential units in a stage are selected. The transition matrix $\boldsymbol{Q}_f$ (renamed for the convenience of this particular example) is as follows:

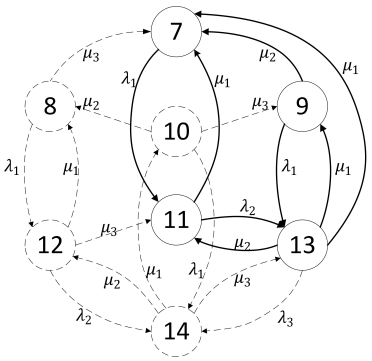
$\boldsymbol{Q}_f =$

| | 7 | 8 | 9 | 10 | 11 | 12 | 13 | **14** |
|---|---|---|---|---|---|---|---|---|
| 7 | $-\lambda_1$ | | | | $\lambda_1$ | | | |
| 8 | $\mu_3$ | $-\lambda_1-\mu_3$ | | | | $\lambda_1$ | | |
| 9 | $\mu_2$ | | $-\lambda_1-\mu_2$ | | | | $\lambda_1$ | |
| 10 | | $\mu_2$ | $\mu_3$ | $-\lambda_1-\mu_2-\mu_3$ | | | | $\boldsymbol{\lambda_1}$ |
| 11 | $\mu_1$ | | | | $-\lambda_2-\mu_1$ | | $\lambda_2$ | |
| 12 | | $\mu_1$ | | | $\mu_3$ | $-\lambda_2-\mu_1-\mu_3$ | | $\boldsymbol{\lambda_2}$ |
| 13 | $\mu_1$ | | $\mu_1$ | | $\mu_2$ | | $-\lambda_3-2\mu_1-\mu_2$ | $\boldsymbol{\lambda_3}$ |
| **14** | | | | $\boldsymbol{\mu_1}$ | | $\boldsymbol{\mu_2}$ | $\boldsymbol{\mu_3}$ | $-\mu_1-\mu_2-\mu_3$ |

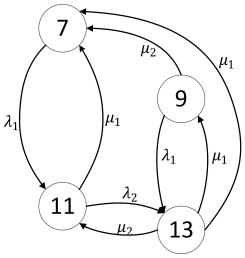
Since the physical meaning of $\lambda_1, \lambda_2, \lambda_3$ and $\mu_1, \mu_2, \mu_3$ are the failure rates and repair rates of potential units in one stage, it is reasonable to assume that $\lambda_1, \lambda_2, \lambda_3$ are of the same order of magnitude, denoted as $O(\lambda)$, and $\mu_1, \mu_2, \mu_3$ of $O(\mu)$. We assume the repair rate to be two orders of magnitudes larger, that is $O(\mu) = 10^2 O(\lambda)$. With that, we examine the order of magnitude of the stationary probability vector of this isolated system by solving the linear system $\{\boldsymbol{\pi}_f^\top \boldsymbol{Q}_f = \boldsymbol{0}, \boldsymbol{\pi}_f^\top \boldsymbol{1} = 1\}$. As seen in Figure 16, the stationary probability of states 8, 10, 12, 14 are of $O(10^{-6})$). Therefore, excluding them should still allow us to account for states whose sum has at least a probability of 0.99996 of all situations. Also, notice that states 7, 9, 11, 13 still consist a closed transitioning loop.

$$\boldsymbol{\pi}_f \sim \begin{pmatrix} \frac{O(\mu)^7}{O(\mu)^7} \\ \frac{O(\lambda)^3 * O(\mu)^4}{9 * O(\mu)^7} \\ \frac{O(\lambda)^2 * O(\mu)^5}{3 * O(\mu)^7} \\ \frac{O(\lambda)^3 * O(\mu)^4}{18 * O(\mu)^7} \\ \frac{O(\lambda)^1 * O(\mu)^6}{O(\mu)^7} \\ \frac{O(\lambda)^3 * O(\mu)^4}{18 * O(\mu)^7} \\ \frac{O(\lambda)^2 * O(\mu)^5}{3 * O(\mu)^7} \\ \frac{O(\lambda)^3 * O(\mu)^4}{9 * O(\mu)^7} \end{pmatrix} = \begin{pmatrix} O(1) \\ \boldsymbol{O(10^{-6})} \\ O(10^{-4}) \\ \boldsymbol{O(10^{-6})} \\ O(10^{-2}) \\ \boldsymbol{O(10^{-6})} \\ O(10^{-4}) \\ \boldsymbol{O(10^{-6})} \end{pmatrix}$$

(a) Order of magnitude estimation




(b) Transition diagram with low probability states




(c) Reduced transition diagram

Figure 16: States with low probabilities are eliminated

## 6. Case studies

The examples in this section are all solved in GAMS 24.8.5 on an Intel® Core™ i7-7700 CPU at 3.60GHz with 4 Cores and 8 Logical Processors.

### *6.1. The impact of incorporating maintenance*

In this section, a small example is solved to show the significance of incorporating maintenance. The superstructure being examined is shown in Figure 17, which has 3 non-identical potential units for stage 1 and 2 non-identical potential units for stage 2. The parameters are shown in Table 3. A time horizon of 10 years is considered with the rate of return as 10%.

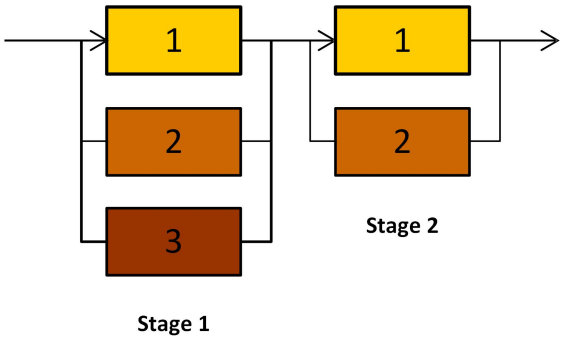



Figure 17: A small example

Table 3: Parameters for the small example

| Stage | Unit | $MTBF$(day) | $MTTR$(day) | Installation cost (k$) | Repair cost rate (k$ per time) |
|---|---|---|---|---|---|
| 1 | 1 | 50 | 7 | 123 | 12 |
| 1 | 2 | 45.5 | 7.7 | 98 | 12 |
| 1 | 3 | 41.7 | 8.3 | 74 | 12 |
| 2 | 1 | 66.7 | 2.6 | 147 | 10 |
| 2 | 2 | 50 | 2.8 | 123 | 10 |

| Revenue rate(k$) | Penalty rate(k$) | Bonus rate(k$) | Availability lower bound | Availability upper bound |
|---|---|---|---|---|
| 700 | 1000 | 1000 | 0.988 | 0.998 |

***Without maintenance consideration***

When considering no maintenance or spare parts, only constraints and equations (23), (26)-(29), (37)-(39), (43), (45)-(47), (62), (C.1)-(C.2) and (C.8)- (C.12) are applied, and the transition matrices are based on $\lambda^0_{k,j}$. Moreover, in the objective function (61), the cost terms $repaCost$, $inspCost$ and $mainCost$ all equal to 0. With that, the model reduces to an MILP, which has 951 equations and 466 variables with 39 binary variables. The MILP model is solved with CPLEX 12.7.1.0 in 0.047 CPUs. The optimal design corresponds to having units 1 and 2 in stage 1, and units 1 and 2 in stage 2, as shown in Figure 18.

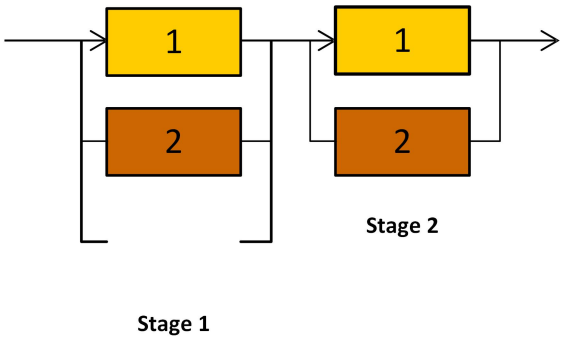



Figure 18: Optimal design considering no maintenance

The expected system availability is 0.989, and the net present value is $2,549,130, with a revenue of $6,922,500 and zero penalty and bonus. $491,600 is spent on unit investment, and $1,974,200 is spent on repair.

***With maintenance consideration***

With the MTBF and MTTR in Table 3 as baseline, and the supplementary parameters shown in Table 4, the problem is solved again considering maintenance and variable failure rates and repair rates ((23), (26)-(29), (37)-(39), (43), (45)-(47), (52)-(60), (62), (C.1)-(C.2) and (C.8)- (C.12)). The options for inspection intervals are 14 days, 30 days (a month), 60 days (two months), 183 days (half a year) and 365 days (a year).

Table 4: Supplementary parameters for maintenance consideration

| Stage | Inspection cost rate (k$ per time) | Maintenance cost rate (k$ per time) | Maintenance time (day) | Deterioration time (day) |
|---|---|---|---|---|
| 1 | 0.1 | 0.6 | 1 | 10 |
| 2 | 0.1 | 0.5 | 1 | 12 |

The MINLP model has 44,938 equations, 44,458 variables with 49 binary variables, and is solved with the global solver BARON 17.4.1 in 638.9 CPUs, and the non-global solver SBB yielding the same result in 62.3 CPUs. For other MINLP solvers, DICOPT reports integer infeasiblity, and ANTIGONE did not solve for 6 hours). The optimal design is to have units 1 and 3 in stage 1, and units 1 and 2 in stage 2, as shown in Figure 19.

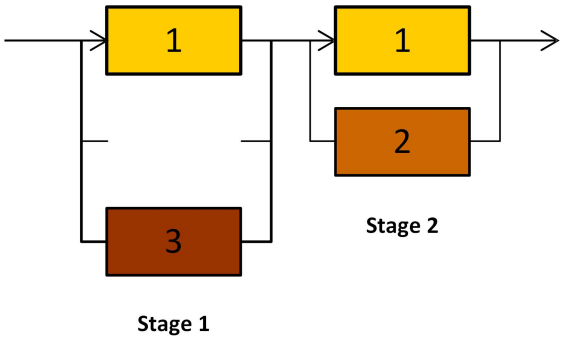



Figure 19: Optimal design considering maintenance

The expected system availability is 0.995. The expected net present value is \$3,269,447, with a revenue of \$6,967,400 and no bonus or penalty. \$467,020 is spent on unit investment, \$52,100 is spent on inspections, \$46,500 is spent on maintenance, and \$788,300 on repair. Other results including inspection intervals, equivalent MTBF's and equivalent MTTR's are shown in Table 5.

Table 5: Optimization results considering maintenance

| Stage | Inspection interval (day) | Equivalent $MTBF$ (day) |
|---|---|---|
| 1 | 14 | Unit2: 98.0; Unit3: 86.2 |
| 2 | 14 | Unit1: 181.8; Unit2: 117.6 |

From the above results we can see that when maintenance is considered, the model suggest additional costs on inspection and maintenance, while spending less on the unit investment and repair, which leads to a higher availability (from 0.989 to 0.995), and a significantly overall higher net present value (from \$2,549,130 to \$3,269,447).

*6.2. Computational performance improvement from model decomposition and scenario reduction*

***Resolve the small problem***

This section reports the result of solving the model using the decomposition scheme and scenario reduction. Like what is shown in section 5.2, 4 states of stage 1 for the design with all 3 units are eliminated, which reduces the number of elements in the system transition matrix by 30%. After one iteration, the problem converges ($\epsilon = 0.1\%$) to the same solution as solving the original problem by BARON. As shown in Figure 20, the optimal design is identical to the design shown in Figure 19.

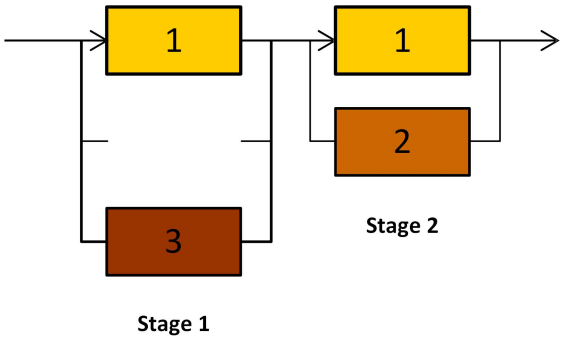



Figure 20: Optimal design by decomposition

The expected system availability is 0.995. The expected net present value is \$3,269,447, with a revenue of \$6,967,400 and no bonus or penalty. \$467,020 is spent on unit investment, \$52,100 is spent on inspections, \$46,400 is spent on maintenance, and \$787,900 on repair. Other results including inspection intervals and equivalent MTBF's are shown in Table 6. Notice here that the small discrepancy between the two sets of results are mainly due to data transferring accuracy issues in the implementation of the decomposition method.

Table 6: Optimization results considering maintenance

| Stage | Inspection interval (day) | Equivalent $MTBF$ (day) |
|---|---|---|
| 1 | 14 | Unit2: 98.0; Unit3: 86.2 |
| 2 | 14 | Unit1: 181.8; Unit2: 117.6 |

More importantly, the computational performance is improved by two orders of magnitude, both in terms of the model size and total CPU time, which is shown in Table 7. The MILP models are solved with CPLEX 12.7.1.0, and the MINLP models are solved with BARON 17.4.1.

Table 7: Computational results of the decomposition method

| | No. Equations | No. Variables | No. Discrete variables | CPUs |
|---|---|---|---|---|
| MILP | 877 | 406 | 39 | 0.235 |
| MINLP | 324 | 329 | 13 | 0.06 |

***A larger example***

Next, the MINLP model based on Markov chain is applied to a larger system, which has 4 stages and 3 non-identical potential units for each stage. The superstructure is shown in Figure 21. Parameters are listed in Table 8, 9 and 10. A time horizon of 10 years is considered with the rate of return as 10%.

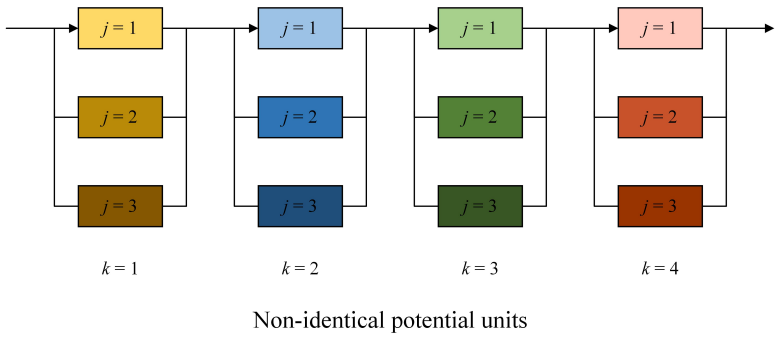



Figure 21: A larger example with 4 stages and 3 non-identical potential units for each stage

Table 8: Reliability and installation cost parameters

| Stage | Unit | $MTBF$(day) | $MTTR$(day) | Installation cost(k$) |
|---|---|---|---|---|
| 1 | 1 | 500 | 7 | 123 |
| | 2 | 455 | 7.7 | 98 |
| | 3 | 417 | 8.3 | 74 |
| 2 | 1 | 667 | 2.6 | 147 |
| | 2 | 500 | 2.8 | 123 |
| | 3 | 450 | 2.9 | 110 |
| 3 | 1 | 625 | 4.2 | 92 |
| | 2 | 588 | 4.3 | 80 |
| | 3 | 556 | 4.5 | 68 |
| 4 | 1 | 1000 | 3.8 | 141 |
| | 2 | 667 | 4.5 | 129 |
| | 3 | 500 | 5.6 | 104 |

Table 9: Repair and maintenance parameters

| Stage | Repair cost rate (k$ per time) | Inspection cost rate (k$ per time) | Maintenance cost rate (k$ per time) | Maintenance time (day) | Inspection window (day) |
|---|---|---|---|---|---|
| 1 | 12 | 0.1 | 1.2 | 1 | 5 |
| 2 | 10 | 0.1 | 1 | 1 | 6 |
| 3 | 10 | 0.1 | 0.8 | 1 | 6 |
| 4 | 12 | 0.1 | 0.6 | 1 | 5 |

Table 10: Profitability parameters

| Revenue rate(k$) | Penalty rate(k$) | Bonus rate(k$) | Availability lower bound | Availability upper bound |
|---|---|---|---|---|
| 700 | 1000 | 1000 | 0.988 | 0.998 |

The original problem is too large to be solved directly since the computer memory would run out to generate the model. Therefore, only the results from using the decomposition scheme and scenario reduction will be presented. Like what is shown in section 5.2, for each stage, 4 states of the design with all 3 units installed are eliminated, which reduces the number of elements in the system transition matrix by 75%. The algorithm converges in 7 iterations ($\epsilon = 1.7\%$) to the flowsheet shown in Figure 22.

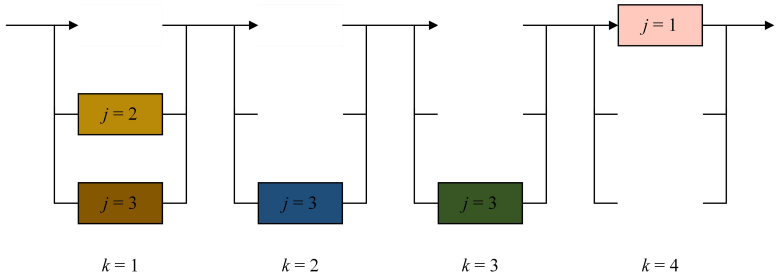



Figure 22: A larger example with 4 stages and 3 non-identical potential units for each stage

The expected system availability is 0.987. The expected net present value is $3,624,198, with a revenue of $6,907,700 and a penalty of $11,900. $491,600 is spent on unit investment, $84,300 is spent on inspections, $6,900 is spent on maintenance, and $106,700 on repair.Other results including inspection intervals and equivalent MTBF's are shown in Table 11.

Table 11: Optimization results considering maintenance

| Stage | Inspection interval (day) | Equivalent $MTBF$ (day) |
|---|---|---|
| 1 | 60 | Unit2: 490; Unit3: 448 |
| 2 | 14 | Unit3: 775 |
| 3 | 14 | Unit3: 952 |
| 4 | 14 | Unit1: 1542 |

Figure 23 shows the iterations of the decomposition method. The optimal solution is actually found at the first iteration. However, the relaxation of the upper bound generating MILP's is not very tight, therefore, the gap remains fairly large.

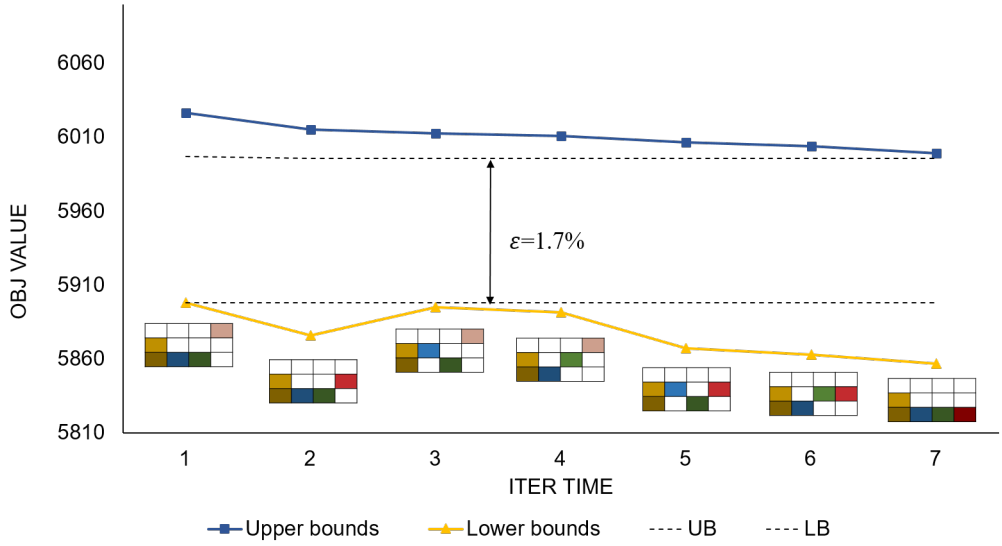



Figure 23: The iteration curve for the 4-stage example

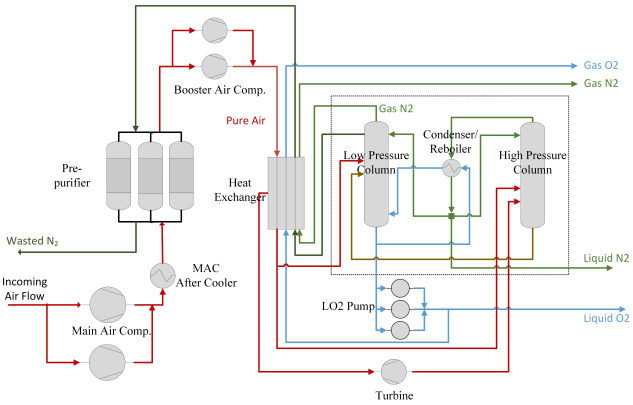
Computational results are shown in Table 12. The MILP models are solved with Xpress 29.01, and the MINLP models are solved with BARON 17.4.1. The total CPU time of the MINLP's is 124.3s, and 1634.5s for the MILP's.

Table 12: Computational results of the 4-stage example

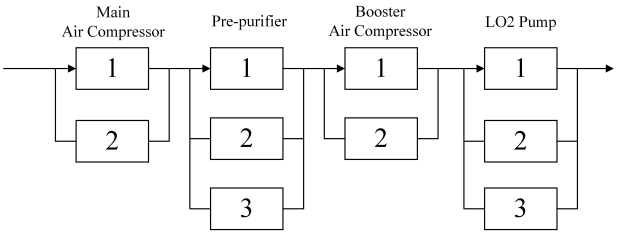
| | MINLP | | | | MILP | | | |
|---|---|---|---|---|---|---|---|---|
| | No. Eqs | No. Vars | No. Disc. Vars | CPUs | No. Eqs | No. Vars | No. Disc. Vars | CPUs |
| 1 | | | | 2.9 | 714294 | | | 274.5 |
| 2 | | | | 2.56 | 714295 | | | 250.3 |
| 3 | | | | 14.8 | 714296 | | | 243.3 |
| 4 | 1108 | 1142 | 23 | 18.4 | 714297 | 236716 | 2444 | 312.4 |
| 5 | | | | 21.3 | 714298 | | | 301.8 |
| 6 | | | | 39.4 | 714299 | | | 265.2 |
| 7 | | | | 24.9 | 714300 | | | 261.5 |
| Total | | | | 124.3 | | | | 1634.5 |

### *6.3. Solve the motivating example: Air separation unit*

Finally, the model is applied to the motivating example of ASU (air separation unit) introduced in section 2, where the compressors have six failure modes and at least 2 units are needed for the pre-purifier. The flowsheet and the superstructure diagram are shown in Figure 24. A time horizon of 10 years is considered.




(a) Typical flowsheet of an air separation unit with potential parallel units




(b) The diagram of ASU reliability design alternatives. Each block represents a parallel unit with certain availability and cost rates

Figure 24: The motivating example-ASU

The $LO_2$ pump stage is a regular stage like the ones in sections 6.1 and 6.2. The main air compressor and the booster air compressor each have 6 failure modes (root causes of failure): rotor, bearing/seals, gearbox, lube oil system, motor, and motor bearing. Appendix D explains how the model changes in the case of multiple failure modes. For the pre-purifier, at least 2 units are required for normal operation, which leads to different state space and transitional relationship than those of the example in section 4.2. The state space where all 3 parallel units are installed for the pre-purifier stage is shown in Table 13 as an example.

Table 13: The state space where all 3 parallel units are installed for the pre-purifier stage

| Design decisions/subspaces | States | unit 1 | unit 2 | unit 3 |
|---|---|---|---|---|
| $T_{k,4}$: unit 1, 2 and 3 | state 13 | active | active | standby |
| | state 14 | active | active | being repaired |
| | state 15 | active | being repaired | active |
| | state 16 | active | being repaired | being repaired |
| | state 17 | being repaired | active | active |
| | state 18 | being repaired | active | being repaired |
| | state 19 | being repaired | being repaired | active |
| | state 20 | being repaired | being repaired | being repaired |

Table 14 shows the failure modes considered for each processing stages. Out of confidentially consideration, the real failure modes are disguised with alias.

Table 14: Failure modes of each processing stage

| Stage | Failure mode | Stage | Failure mode | Stage | Failure mode | Stage | Failure mode |
|---|---|---|---|---|---|---|---|
| Main air compressor | FMC1<br>FMC2<br>FMC3<br>FMC4<br>FMC5<br>FMC6 | Pre-purifier | FMPF1 | Booster air compressor | FMC1<br>FMC2<br>FMC3<br>FMC4<br>FMC5<br>FMC6 | $LO_2$ Pump | FMP1 |

Relevant cost and reliability parameters of the units need to be concealed as well. Mean time between failures (MTBF) range from 5-25 years. Mean time to repair (MTTR) range from 8 - 1080 hours. Capital cost of each unit range from \$85k - \$800k. Repair costs range from \$2k - \$20k per time. Inspection costs range from \$0.05k - \$0.5k per time. Maintenance costs range from \$1k - \$10k per time. Maintenance times range between 1 and 2 days. Inspection window lengths range between 5 and 6 day. Table 15 shows the profitability parameters used in the model.

Table 15: Profitability parameters

| Revenue rate(k\$) | Penalty rate(k\$) | Bonus rate(k\$) | Availability lower bound | Availability upper bound |
|---|---|---|---|---|
| 3000 | 30000 | 1000 | 0.988 | 0.998 |

In order to reduce the model size, low probability states generated from the two-unit design of the MAC and the BAC, and the three-unit design of the pre-purifier stage and the $LO_2$ pump stage are eliminated. The reduce state space for the MAC and the BAC are shown in Table 16, and the pre-purifier in Table 17. The $LO_2$ pump stage is the same as the example described in section 5.2 and will not be reiterated.

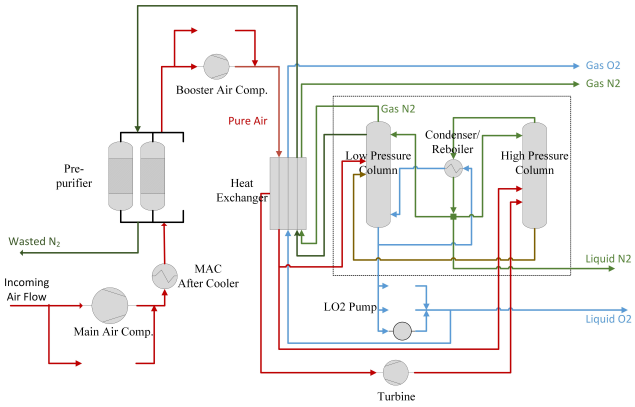
Table 16: The reduced state space generated from the two-unit design of the MAC and BAC stages

| Design decisions/subspaces | States | unit 1 | unit 2 |
|---|---|---|---|
| unit 1 and 2 | state 15 | active | standby |
| | state 16 | active | FMC6 |
| | state 17 | FMC1 | active |
| | state 18 | FMC2 | active |
| | state 19 | FMC3 | active |
| | state 20 | FMC4 | active |
| | state 21 | FMC5 | active |
| | state 22 | FMC6 | active |
| | state 23 | FMC6 | FMC6 |

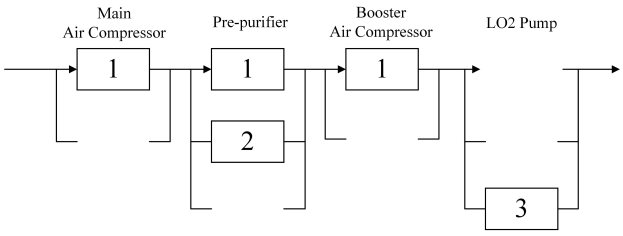
Table 17: The reduced state space generated from the three-unit design of the pre-purifier stage

| Design decisions/subspaces | States | unit 1 | unit 2 | unit 3 |
|---|---|---|---|---|
| unit 1, 2 and 3 | state 13 | active | active | standby |
| | state 14 | active | active | being repaired |
| | state 15 | active | being repaired | active |
| | state 16 | active | being repaired | being repaired |
| | state 17 | being repaired | active | active |

Again, the original model is too large to be solved directly. Therefore, only the results from using the decomposition scheme and scenario reduction are presented. The algorithm converges in 7 rounds ($\epsilon = 1.7\%$) to the flowsheet shown in Figure 25. Only the least number of units are selected for each stage. For the main air compressor and the booster air compressor, more reliable and expensive units are selected, while for the $LO_2$ pump, the solution goes for the cheapest one.




(a) The flowsheet representation




(b) The block diagram representation

Figure 25: The optimal design of the ASU example

The expected system availability is 0.9866. The expected net present value is \$15649.4k, with a revenue of \$29597.8k and a penalty of \$421.9k. \$2083k is spent on unit investment, \$262.7k is spent on inspections, \$7.4k is spent on maintenance, and \$48.9k on repair. The inspection intervals are shown in Table 18. Qualitatively speaking, this solution tends to spend more efforts on reducing the failure rates for failure modes with longer repair time.

Table 18: Inspection interval decisions

| Stage | Unit | Failure mode | Inspection interval (day) |
|---|---|---|---|
| Main air compressor | 1 | FMC1 | 14 |
| | | FMC2 | 14 |
| | | FMC3 | 14 |
| | | FMC4 | 365 |
| | | FMC5 | 14 |
| | | FMC6 | 14 |
| Pre-purifier | 1<br>2 | FMPF1 | 14 |
| Booster air compressor | 1 | FMC1 | 14 |
| | | FMC2 | 14 |
| | | FMC3 | 14 |
| | | FMC4 | 365 |
| | | FMC5 | 14 |
| | | FMC6 | 14 |
| $LO_3$ Pump | 2 | FMP1 | 365 |

Figure 26 shows the iteration process, where the optimum is found at the second solve.

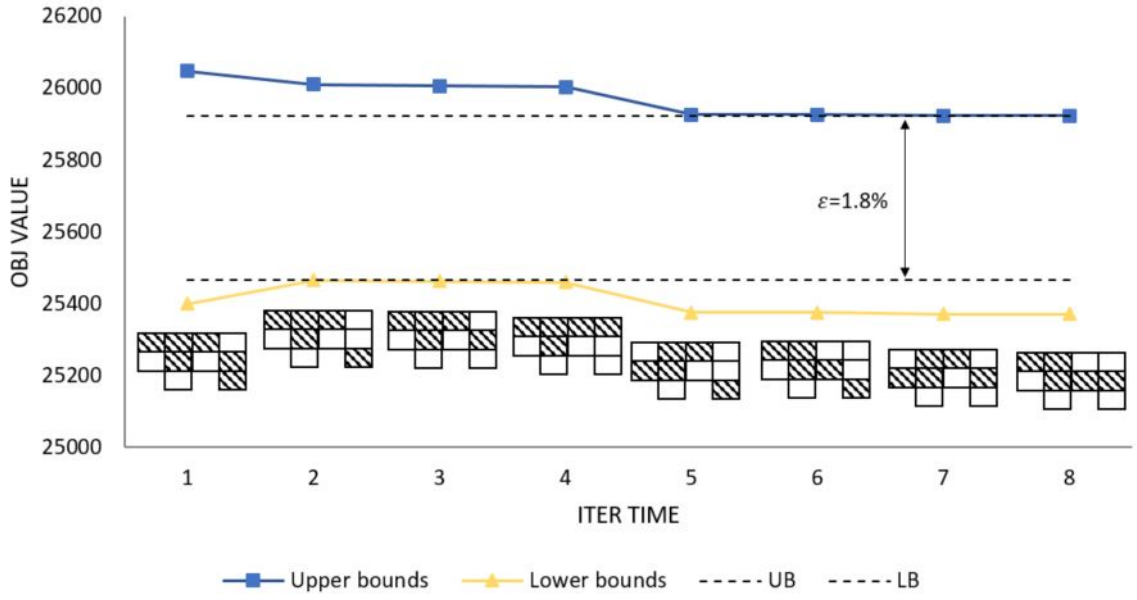


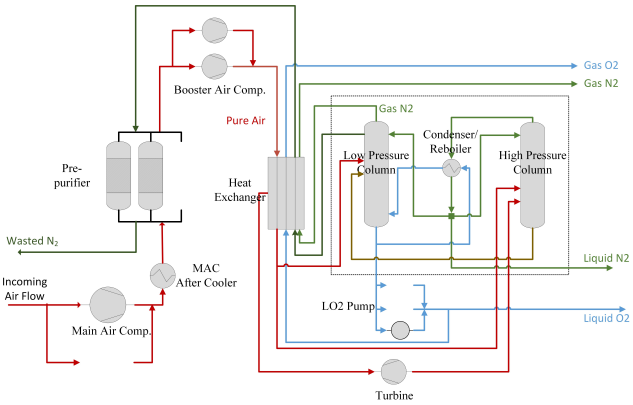

Figure 26: The iteration curve for the 4-stage example

Computational results are shown in Table 19. The MILP models are solved with Xpress 29.01, and the MINLP models are solved with SBB 25.1.1. The total CPU time of the MINLPs is 492.01s, and 748.82s for the MILPs. Note that arguably, the optimal solution cannot be guaranteed as SBB is not a global solver.

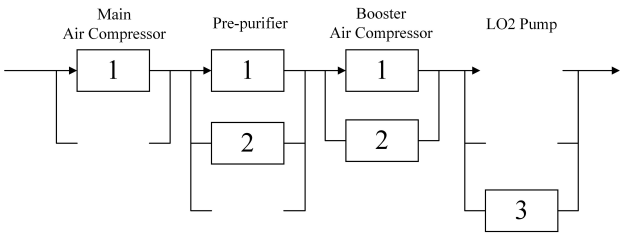
Table 19: Computational results of the 4-stage example

| | MINLP | | | | MILP | | | |
|---|---|---|---|---|---|---|---|---|
| | No. Eqs | No. Vars | No. Disc. Vars | CPUs | No. Eqs | No. Vars | No. Disc. Vars | CPUs |
| 1 | 346524 | 346577 | | 511.9 | 490151 | | | 100.66 |
| 2 | 86911 | 86965 | | 60.14 | 490153 | | | 130.30 |
| 3 | 86911 | 86965 | | 86.39 | 490154 | | | 99.85 |
| 4 | 86911 | 86965 | 73 | 72.69 | 490155 | 163230 | 282 | 103.59 |
| 5 | 86911 | 86965 | | 72.04 | 490156 | | | 115.92 |
| 6 | 86911 | 86965 | | 62.84 | 490157 | | | 88.47 |
| 7 | 86911 | 86965 | | 60.59 | 490158 | | | 110.03 |

In addition, we look into an extreme case where the availability lower bound is pushed from 0.988 up to 0.995, and the optimal design is as shown in Figure 27. The expected system availability is 0.9925. The expected net present value is \$15017.7k, with a revenue of \$29776.3k and a penalty of \$737.3k. \$2697.7k is spent on unit investment, \$200.0k is spent on inspections, \$4.0k is spent on maintenance, and \$56.1k on repair.




(a) The flowsheet representation




(b) The block diagram representation

Figure 27: The optimal design of the ASU example

Comparing to the original case, the second booster air compressor is added to the flowsheet. Regarding this change, Table 20 lists the lump sum probability of failure scenarios involving each processing stage for the normal case and the extreme case. It can be seen that both MAC and

BAC are the weak points of the first design. When facing extremely high availability requirement, a unit from the cheaper stage, BAC is added, leading to a sharp decrease of its failure time.

Table 20: Probability distribution among failure scenarios

| | Availability lower bound | 0.988 | 0.995 |
|---|---|---|---|
| | Actual availability | 0.9866 | 0.9925 |
| Probabilities | MAC | 0.006 | 0.006 |
| | PPF | 0.001 | 0.001 |
| | BAC | 0.006 | $1.27*10^{-5}$ |
| | LO2 pump | $2.34*10^{-4}$ | $2.34*10^{-4}$ |

## 7. Conclusions

This paper extends our recent work (Ye et al., in press) which proposes a general mixed-integer framework for the optimal selection of redundancy bearing reliability concerns based on fixed probability of failure for single potential units. In this paper, the stochastic process of system failures and repairs is modeled as a continuous-time Markov chain, moreover, the impact of maintenance is incorporated.

With a general air separation unit as the motivating example, two strategies are considered to increase the availability of the system. The first strategy is to install parallel units for certain processing stages, such that when the primary unit fails, the other units can fill in its place in order to reduce system downtime. The second strategy is to carry out periodic inspections, and follow-up maintenance if the inspection results indicate that the equipment will fail shortly. This strategy allows the system to avoid a number of unplanned shutdowns and repairs, which are much more costly than planned maintenance in terms of both time and money. A non-convex MINLP model is proposed accordingly. When the inspection frequencies are fixed, the model reduces to an MILP. A small superstructure with two stages is solved directly with global solvers to show that disregarding maintenance can reduce computing time but yields inferior solutions in terms of system availability and net present value.

The non-convex MINLP model does not scale well, and has a large number of bilinear and multilinear terms. In order to overcome these computational difficulties, a decomposition scheme is proposed to reduce the size of the model and the computational time. Moreover, scenario reduction, i. e., pre-process the scenarios and eliminate those with consistently low probabilities, is also applied to reduce the model scale. The specialized solution method is applied to the two-stage problem and reaches global optimum in orders of magnitude less time than the global solvers. The method is also successfully applied to a system that is too large to be solved directly, which has 4 stages and 3 non-identical units for each stage. Finally, the motivating example of the air separation unit is solved with the proposed specialized solution method.

## 8. Acknowledgment

The authors would like to acknowledge the support from National Science Foundation under grant CBET-1705372, Praxair, and Center for Advanced Process Decision-making at Carnegie Mellon University.

## 9. Nomenclature

Index

| | |
|---|---|
| $k$ | stages |
| $l$ | options of inspection intervals |
| $j$ | Units |
| $s$ | States of single stages |
| $r$ | Alias of $s$ |
| $\bar{s}$ | States of the entire system |
| $\bar{r}$ | Alias of $\bar{s}$ |
| $h$ | Designs of single stages |
| $\bar{h}$ | Designs of the entire system |

Set

| | |
|---|---|
| $K$ | Set of all stages |
| $J_k$ | Set of potential units of stage $k$ |
| $S_k$ | super state space of stage $k$ |
| $\bar{S}$ | super state space for the entire system |
| $H_k$ | Set of potential designs of stage $k$ |
| $\bar{H}$ | Set of potential designs of the entire system |
| $\bar{S}$ | Set of system states |
| $\bar{S}^f$ | Set of system states that are failed |
| $SC$ | Set of the index tuples where $s$ in $S_k$ corresponds to $\bar{s}$ in $\bar{S}$ |
| $HC$ | Set of the index tuples where $h$ in $H_k$ corresponds to $\bar{h}$ in $H_k$ |

Parameter

| | |
|---|---|
| $\mu_{k,j}$ | Repair rate of unit $j$ in stage $k$ |
| $\boldsymbol{Q}_k$ | super transition rate matrix of stage $k$ |
| $\boldsymbol{W}$ | super transition rate matrix of the entire system |
| $c\_inst_{k,j}$ | Investment cost of single unit $j$ in stage $k$ |
| $c\_repa_k$ | Repair cost of single units in stage $k$ |
| $rv$ | Revenue rate of final products |
| $pn$ | Penalty rate for not meeting lower bound of availability |
| $bn$ | Bonus rate for exceeding upper bound of availability |
| $A\_lo$ | The lower bound of system availability arranged in the contract |
| $A\_up$ | The upper bound of system availability arranged in the contract |
| $\lambda^0_{k,j}$ | Failure rate of unit $j$ in stage $k$ without maintenance |
| $T_l^{insp}$ | Options of inspection interval |
| $T_k^d$ | Deterioration time of equipment in stage $k$ |
| $c\_insp_k$ | Inspection cost (per time) of single units in stage $k$ |
| $c\_main_k$ | Maintenance cost (per time) of single units in stage $k$ |
| $c\_repa_k$ | Repair cost (per time) of single units in stage $k$ |
| $T\_main_k$ | Downtime of single maintenance of stage $k$ |
| $\tau$ | Number of years in the time horizon |
| $r$ | Return rate of cash flow |

Variable

| | |
|---|---|
| $x^{k,l}$ | Binary variable that indictate whether an inspection interval of stage $k$ is selected |
| $t_k^{insp}$ | Inspection interval of stage $k$ |
| $\lambda_{k,j}$ | Failure rate of the units in stage $k$ |
| $y_{k,j}$ | Binary variable that indicate whether to install unit j of stage k |
| $z_{k,h}$ | Binary variable that indicate whether design $h$ in stage $k$ is selected |
| $zz_{k,s}$ | Binary variable that indicate whether state $s$ of stage $k$ exist |
| $\bar{z}_{\bar{h}}$ | Binary variable that indicate whether design $\bar{h}$ is selected |
| $\overline{zz}_{\bar{s}}$ | Binary variable that indicate whether state $\bar{s}$ exist (modeled only as positive variable with bounds) |
| $\pi_{\bar{s}}$ | Stationary probability of state $\bar{s}$ |
| $inspCost$ | Inspection cost |
| $mainCost$ | Maintenance cost |
| $repaCost$ | Repair cost |
| $mainTime$ | Maintenance downtime |
| $A$ | Availability of the system |
| $A^{net}$ | Availability of the system considering downtime caused by maintenance |
| $RV$ | Expected revenue |
| $PN$ | Expected penalty |
| $BN$ | Expected bonus |
| $NPV$ | Net Present Value |
| $w_1, w_2, w_3$ | Binary variables that indicate the range that $A$ lies in |
| $A^1, A^2, A^3$ | Components of the system availability $A$ for corresponding ranges |
| $PN^1, PN^2, PN^3$ | Components of the expected penalty $PN$ for corresponding ranges |
| $BN^1, BN^2, BN^3$ | Components of the expected bonus $BN$ for corresponding ranges |

## Appendix A Continuous-time Markov chain (CTMC)

Continuous-time Markov chain theory has been used in plenty of previous works to describe systems where the failures/repairs of single units are subject to independent stochastic processes. In this section, a formal definition of a continuous-time Markov chain (CTMC) based on Bayesian statistics is given.

For a system that transitions randomly among a finite set of states, if the time it spends in one state follows an exponential distribution, this state-transitioning process is a continuous-time Markov chain. An important property of the Markov chain, also called the Markov property, is that future behaviors of the system depend only on the current state of the model, but not on its historical behavior. Specifically, let $S = \{1, 2, \dots, \}$ be the state space of a random process $X = \{X(t), t \geq 0\}$, which means that $X$ will be varying among the states included in $S$, as shown in Figure A.1. If for any time points $0 \leq T_1 \leq T_2 \leq \cdots \leq T_{n+1}$, and states $i_1, i_2, \dots, i_{n+1} \in S$, there is equation (A.1)

$$Pr\{X(T_{n+1}) = i_{n+1} | X(0) = i_0, X(T_1) = i_1, \dots, X(T_n) = i_n\} = Pr\{X(T_{n+1}) = i_{n+1} | X(T_n) = i_n\} \tag{A.1}$$

then the random process $X = \{X(t), t \geq 0\}$ is called a continuous-time Markov process. Furthermore, if for any $s, t \geq 0, i, j \in S$, there is equation (A.2)

$$Pr\{X(s+t) = j | X(s) = i\} = Pr\{X(t) = j | X(0) = i\} \tag{A.2}$$

then $X = \{X(t), t \geq 0\}$ is time-homogeneous (i.e. the transitional behavior does not change with time), which we are going to focus on.

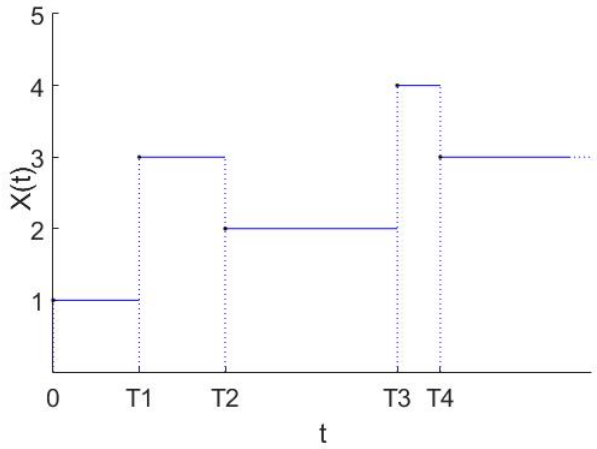



Figure A.1: An example for continuous-time Markov chain

## Appendix B Extended continuous-time Markov chain

The superstructure discuss in this work contains several of the small system shown in section 4.2, for which the time and efforts needed to enumerate over the entire state space and to construct the $\boldsymbol{Q}$ matrix grow geometrically. Actually, the CTMC of the entire system can be formulated based on the CTMC of each single stage, which only requires duplication of what is done in section 4.2. The system CTMC is called the extended CTMC of stage CTMC's.

### *B.1 CTMC and Phase-type distribution (PH-distribution)*

In this section we introduce a class of probability distribution that is closely connected to CTMC, phase-type distribution (Neuts, 1981). The concept of PH-distribution and its applicability in constructing extended Markov process are fundamentally important to the model formulation in this article.

Given a CTMC on the state space $\{1, \ldots, m+1\}$ with generator

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{T}^0 \\ \boldsymbol{t} & t^0 \end{bmatrix} \tag{B.1}$$

where $\boldsymbol{T}$ is of dimension $m \times m$, and $\boldsymbol{T}^0$ is of dimension $m \times 1$.According to the properties of the generator stated in section 3.3, we have

$$\boldsymbol{Te} + \boldsymbol{T^0} = \boldsymbol{0} \tag{B.2}$$

which means that $\boldsymbol{T^0}$ is known once $\boldsymbol{T}$ is determined.
If the initial probability(the probability of a state to be the starting state) vector is given by $[\boldsymbol{\alpha}, \alpha_{m+1}]$, we say that the probability distribution $F(\cdot)$ of the time until first reaching state $m+1$ follows the PH-distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order $m$. The probability distribution is given by

$$F(t) = 1 - \boldsymbol{\alpha} exp(\boldsymbol{T}t)\boldsymbol{e} \tag{B.3}$$

Note that the initial probability distribution has to satisfy that

$$\boldsymbol{\alpha e} + \alpha_{m+1} = 1 \tag{B.4}$$

Thus, $\boldsymbol{\alpha}$ gives full information of the initial state. Consider the example in section 3.3, given the initial probability vector $[\boldsymbol{\alpha}, \alpha_4]$ of the 4 states, the probability distribution $F(\cdot)$ of the time until reaching state 4 follows the PH-distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order 3, where

$$\boldsymbol{T} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} -\lambda_1 - \lambda_2 & \lambda_1 & \lambda_2 \\ \mu_1 & -\mu_1 - \lambda_2 & 0 \\ \mu_2 & 0 & -\mu_2 - \lambda_1 \end{bmatrix} \end{array} \tag{B.5}$$

Also, the fourth column of $\boldsymbol{Q}$ in (B.1) with the fourth row suppressed is $\boldsymbol{T^0}$, which indicates the transition rate from the first 3 states to the broken state.

$$\boldsymbol{T^0} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{c} 4 \end{array} \\ \begin{bmatrix} 0 \\ \lambda_2 \\ \lambda_1 \end{bmatrix} \end{array} \tag{B.6}$$

*B.2 From PH-distribution to extended CTMC*

Let $X$ and $Y$ be independent random variables with PH-distributions $F(\cdot)$ and $G(\cdot)$ as shown in Table B.1.

Table B.1: Information about $X$ and $Y$

| Random variable | cdf. | Representation | Order |
|---|---|---|---|
| $X$ | $F(\cdot)$ | $(\boldsymbol{\alpha}, \boldsymbol{T})$ | $m$ |
| $Y$ | $G(\cdot)$ | $(\boldsymbol{\beta}, \boldsymbol{S})$ | $n$ |

Let $H(\cdot)$ be the probability distribution of $Z = max(X, Y)$, then $H(\cdot)$ is also a PH-distribution. $H(\cdot)$ has the representation $(\boldsymbol{\gamma}, \boldsymbol{L})$ of order $mn + m + n$, given by (Neuts, 1981)

$$\boldsymbol{\gamma} = [\boldsymbol{\alpha} \otimes \boldsymbol{\beta}, \beta_{n+1}\boldsymbol{\alpha}, \alpha_{m+1}\boldsymbol{\beta}] \tag{B.7}$$

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{S} & \boldsymbol{I} \otimes \boldsymbol{S}^0 & \boldsymbol{T}^0 \otimes \boldsymbol{I} \\ 0 & \boldsymbol{T} & 0 \\ 0 & 0 & \boldsymbol{S} \end{bmatrix} \tag{B.8}$$

It is evident that the representation of $\boldsymbol{H}$ can be adapted to a transition rate matrix that describes the behavior of a system with two subsystems whose time-to-break(the time from starting point to the first break down) follows independent phase-type distributions $F(\cdot)$ and $G(\cdot)$
First we manipulate $\boldsymbol{L}$ into a more compact form:
(1) Exchange the 2nd and the 3rd column blocks and the 2nd and the 3rd row blocks respectively.

$$\boldsymbol{L'} = \begin{bmatrix} \boldsymbol{I}_{mn\times mn} & 0 & 0 \\ 0 & 0 & \boldsymbol{I}_{n\times n} \\ 0 & \boldsymbol{I}_{m\times m} & 0 \end{bmatrix} \boldsymbol{L} \begin{bmatrix} \boldsymbol{I}_{mn\times mn} & 0 & 0 \\ 0 & 0 & \boldsymbol{I}_{m\times m} \\ 0 & \boldsymbol{I}_{n\times n} & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{S} & \boldsymbol{T^0}\otimes\boldsymbol{I} & \boldsymbol{I}\otimes\boldsymbol{S^0} \\ 0 & \boldsymbol{S} & 0 \\ 0 & 0 & \boldsymbol{T} \end{bmatrix} \tag{B.9}$$

(2) Split $\boldsymbol{L'}$ into $\boldsymbol{T}$ part and $\boldsymbol{S}$ part.

$$\begin{aligned} \boldsymbol{L'} &= \begin{bmatrix} \boldsymbol{T}\otimes\boldsymbol{I} & \boldsymbol{T^0}\otimes\boldsymbol{I} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{T} \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}\otimes\boldsymbol{S} & 0 & \boldsymbol{I}\otimes\boldsymbol{S^0} \\ 0 & \boldsymbol{S} & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} \boldsymbol{T} & \boldsymbol{T^0} \\ 0 & 0 \end{bmatrix} \otimes \boldsymbol{I}_{n\times n} & 0 \\ 0 & \boldsymbol{T} \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_{(m+1)\times(m+1)} \otimes \boldsymbol{S} & \begin{bmatrix} \boldsymbol{I}_{m\times m}\otimes\boldsymbol{S^0} \\ 0_{n\times m} \end{bmatrix} \\ 0 & 0 \end{bmatrix} \end{aligned} \tag{B.10}$$

(3) Do row and column operations to bring $[\boldsymbol{T}, \boldsymbol{T^0}]$ and $[\boldsymbol{S}, \boldsymbol{S^0}]$ together respectively:
Let the manipulating matrices $\boldsymbol{V}$ and $\boldsymbol{U}$ be

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{I}_{m\times m}\otimes\begin{bmatrix} I_{n\times n} & 0_{n\times 1} \end{bmatrix} & & 0_{mn\times n} \\ 0_{n\times m(n+1)} & & \boldsymbol{I}_{n\times n} \\ \begin{bmatrix} 0_{m\times n} & \boldsymbol{e^1} & 0_{m\times n} & \boldsymbol{e^2} & \cdots & 0_{m\times n} & \boldsymbol{e^m} \end{bmatrix} & & 0_{m\times n} \end{bmatrix} \tag{B.11}$$

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{I}_{m\times m}\otimes\begin{bmatrix} \boldsymbol{I}_{n\times n} \\ 0_{1\times n} \end{bmatrix} & 0_{m(n+1)\times n} & \begin{bmatrix} 0_{n\times m} \\ \boldsymbol{e^1} \\ 0_{n\times m} \\ \boldsymbol{e^2} \\ \vdots \\ 0_{n\times m} \\ \boldsymbol{e^m} \end{bmatrix} \\ 0_{n\times mn} & \boldsymbol{I}_{n\times n} & 0_{n\times m} \end{bmatrix} \tag{B.12}$$

Then

$$\boldsymbol{L''} = \boldsymbol{U}\boldsymbol{L'}\boldsymbol{V} = \begin{bmatrix} \boldsymbol{T}\otimes\boldsymbol{I}_{(n+1)\times(n+1)} & \boldsymbol{T^0}\otimes\begin{bmatrix} \boldsymbol{I}_{n\times n} \\ 0_{1\times n} \end{bmatrix} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_{m\times m}\otimes\begin{bmatrix} \boldsymbol{S} & \boldsymbol{S^0} \\ 0 & 0 \end{bmatrix} & 0 \\ 0 & \boldsymbol{S} \end{bmatrix} \tag{B.13}$$

According to (B.2),

$$\begin{bmatrix} \boldsymbol{L''} & L^0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \boldsymbol{T} & \boldsymbol{T^0} \end{bmatrix} \otimes \boldsymbol{I}_{(n+1)\times(n+1)} \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_{m\times m}\otimes\begin{bmatrix} \boldsymbol{S} & \boldsymbol{S^0} \\ 0 & 0 \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} \boldsymbol{S} & \boldsymbol{S^0} \end{bmatrix} \end{bmatrix} \tag{B.14}$$

Then

$$\begin{bmatrix} \boldsymbol{L''} & L^0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{T^0} \\ 0 & 0 \end{bmatrix} \otimes \boldsymbol{I}_{(n+1)\times(n+1)} + \boldsymbol{I}_{(m+1)\times(m+1)} \otimes \begin{bmatrix} \boldsymbol{S} & \boldsymbol{S^0} \\ 0 & 0 \end{bmatrix} \tag{B.15}$$

Note that both $\begin{bmatrix} \boldsymbol{T} & \boldsymbol{T^0} \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{S} & \boldsymbol{S^0} \\ 0 & 0 \end{bmatrix}$ are the generator of CTMC's. We can then come to the conclusion that, a system with two independent elements that following independent CTMC's generated by $\boldsymbol{Q_1}$ and $\boldsymbol{Q_2}$ follows an extended CTMC, which is generated by $\boldsymbol{Q} = \boldsymbol{Q_1} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{Q_2}$.

## Appendix C Revenue, penalty and bonus

The total revenue is considered proportional to the availability of the system.

$$RV = rvA^{net} \tag{C.1}$$

Generally, in the contract between the plant and the customer, two reference bounds will be set for the availability of the plant (Ye et al., 2017). As shown in Figure (C.1), if the actual availability of the plant does not meet the lower bound, the plant that provides products for the customer will be charged a penalty proportional to the difference. On the other hand, to encourage the plant to increase its availability, if the actual availability exceeds the upper bound, the customer will reward the plant with bonus that is also proportional to the difference.

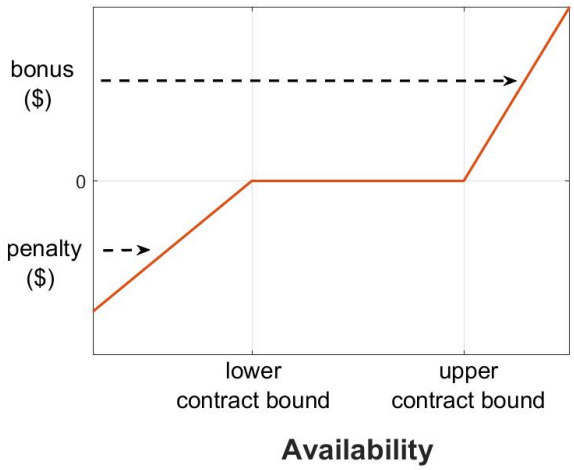



Figure C.1: Definition of penalty and bonus functions

Thus, the penalty and the bonus are described by the equation (C.2) and the disjunction (C.3).

$$w_1 + w_2 + w_3 = 1, \quad w_1, w_2, w_3 = \{0, 1\} \tag{C.2}$$

$$\begin{pmatrix} w_1 = 1 \\ A^{net} \leq A\_lo \\ PN = (A\_lo - A)pn \\ BN = 0 \end{pmatrix} \vee \begin{pmatrix} w_2 = 1 \\ A\_lo \leq A^{net} \leq A\_up \\ PN = 0 \\ BN = 0 \end{pmatrix} \vee \begin{pmatrix} w_3 = 1 \\ A^{net} \geq A\_up \\ PN = 0 \\ BN = (A^{net} - A\_up)bn \end{pmatrix} \tag{C.3}$$

Applying the convex-hull reformulation (Balas, 1985), let

$$A^{net} = A^1 + A^2 + A^3 \tag{C.4}$$

$$PN = PN^1 + PN^2 + PN^3 \tag{C.5}$$

$$BN = BN^1 + BN^2 + BN^3 \tag{C.6}$$

$$\begin{pmatrix} w_1 = 1 \\ A^1 \leq A\_lo \\ PN^1 = (A\_lo - A^1)pn \\ BN = 0 \end{pmatrix} \vee \begin{pmatrix} w_2 = 1 \\ A\_lo \leq A^2 \leq A\_up \\ PN^2 = 0 \\ BN^2 = 0 \end{pmatrix} \vee \begin{pmatrix} w_3 = 1 \\ A^3 \geq A\_up \\ PN^3 = 0 \\ BN^3 = (A^3 - A\_up)bn \end{pmatrix} \tag{C.7}$$

Then, the linear relaxation of equations (C.2) and (C.8) - (C.12) gives the convex hull of (C.2) and (C.3)

$$PN = PN^1 = (w_1 A\_lo - A^1)pn \tag{C.8}$$

$$BN = BN^3 = (A^3 - w_3 A\_up)bn \tag{C.9}$$

$$A^1 \leq w_1 A\_lo \tag{C.10}$$

$$w_2 A\_lo \leq A^2 \leq w_2 A\_up \tag{C.11}$$

$$A^3 \leq w_3 A\_up \tag{C.12}$$

## Appendix D Extension to the circumstance of multiple failure modes

In practice, it is common for a processing unit to fail in not just one way. For example, a compressor can fail because of the deterioration of its motor, rotor, or gearbox, etc., and each of these failures can have different failure rates and repair rates. Therefore, it is necessary to consider multiple failure modes for this kind of units. In this appendix we show how multiple failure modes can be captured by the transition matrix with a small example and how the model can be slightly adapted to accommodate this extension. Consider the example stage shown in Figure D.1 with two potential units.

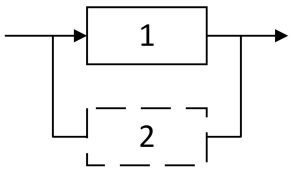



Figure D.1: A single stage $k$

Instead of the simple "being repaired", now the unit is considered to have two independent failure modes, "failure A" and "failure B". The state space considering single failure mode and multiple failure modes are respectively shown in Table D.1 and Table D.2.

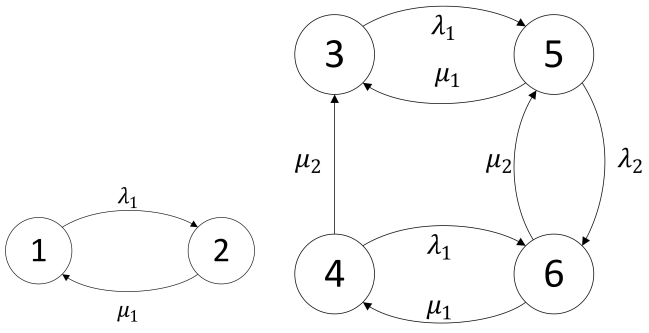
Table D.1: State enumeration for a stage with identical redundancies

| Design decisions/subspace | States | unit 1 | unit 2 |
|---|---|---|---|
| $T_{k,1}$: unit 1 | state 1 | active | |
| | state 2 | being repaired | |
| $T_{k,2}$: unit 1 and 2 | state 3 | active | standby |
| | state 4 | active | being repaired |
| | state 5 | being repaired | active |
| | state 6 | being repaired | being repaired |

Table D.2: State enumeration for a stage with identical redundancies

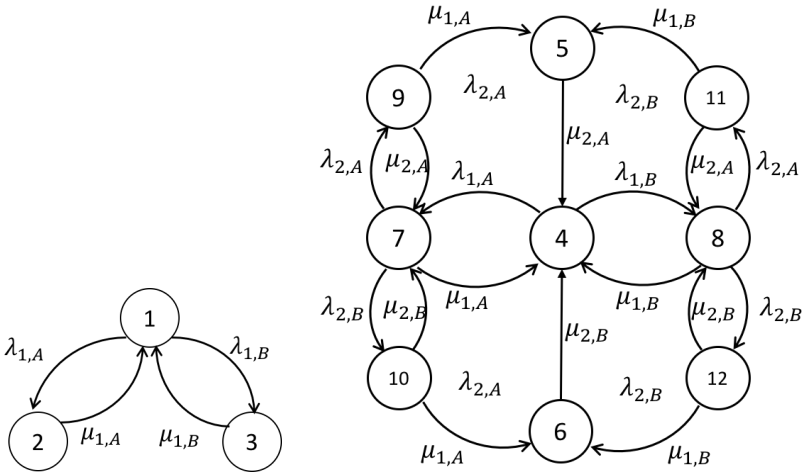
| Design decisions/subspace | States | unit 1 | unit 2 |
|---|---|---|---|
| $T_{k,1}$: unit 1 | state 1 | active | |
| | state 2 | failure A | |
| | state 3 | failure B | |
| $T_{k,2}$: unit 1 and 2 | state 4 | active | standby |
| | state 5 | active | failure A |
| | state 6 | active | failure B |
| | state 7 | failure A | active |
| | state 8 | failure B | active |
| | state 9 | failure A | failure A |
| | state 10 | failure A | failure B |
| | state 11 | failure B | failure A |
| | state 12 | failure B | failure B |

Figure D.2 and Figure D.3 show the corresponding state transition diagrams for the example stage discussed in section 4.2




(a) subspace1($T_{k,1}$) (b) subspace2($T_{k,2}$)

Figure D.2: State transition diagram with single failure mode




(a) subspace1($T_{k,1}$) (b) subspace2($T_{k,2}$)

Figure D.3: State transition diagram with two failure modes

It can be seen from Figure D.2 and Figure D.3 that compared to single failure mode, having multiple failure modes adds complexity to constructing the state space and transition matrices of single stages, which however, is independent with the connections between single stages and the entire system, and how the system availability is represented. Therefore, the model equations and inequalities remain largely unchanged except for the correlation between maintenance and failure rates described in section 4.5:

- A new index is added: $f$ for failure modes. New sets are added: $F_k$ is the set of failure modes of stage $k$. $FAILURE_{\bar{s}}$ is redefined as the set of tuples of $(k, j, f)$ where unit $j$ of stage $k$ is in failure mode $f$ at state $\bar{s}$.

- As is reflected in Figure D.3, to distinguish the failure rates and repair rates between different failure modes of single units, the parameters $\lambda^0_{k,j}$, $\mu^0_{k,j}$ and variable $\lambda_{k,j}$ are rewritten as $\lambda^0_{k,j,f}$, $\mu^0_{k,j,f}$ and $\lambda_{k,j,f}$, respectively.

- Inspections are failure modes specific. Therefore, parameter $T^{insp}_l$, $T^d_k$ and $c\_insp_k$ are rewritten as $T^{insp}_{l,f}$, $T^d_{k,f}$ and $c\_insp_{k,f}$. An inspection interval $t^{insp}_k$ is determined for each processing stage $k$ and each failure mode $f$ that stage $k$ has, hence variables $t^{insp}_k$, $x_{k,l}$ are rewritten as $t^{insp}_{k,f}$ and $x_{k,f,l}$, respectively. Accordingly, equations (52), (53), (54) and (55) are modified as (D.1), (D.2), (D.3) and (D.4) in order to detail down to failure mode level.

$$\sum_{l\in L} x_{k,f,l} = 1, \quad \forall k \in K, f \in F_k \tag{D.1}$$

$$t^{insp}_{k,f} = \sum_{l\in L} x_{k,f,l} T^{insp}_{l,f}, \quad \forall k \in K, f \in F_k \tag{D.2}$$

$$\lambda^0_{k,j,f} - \lambda_{k,j,f} = \sum_{l\in L} x_{k,f,l}(e^{-\lambda^0_{k,j,f}T^{insp}_{l,f}} - e^{-\lambda^0_{k,j,f}(T^{insp}_{l,f}+T^d_{k,f})})/T^{insp}_{l,f}, \quad \forall k \in K, j \in J_k, f \in F_k$$

(D.3)

$$inspCost = \sum_{k \in K} \sum_{f \in F_k} (c\_insp_{k,f} \sum_{l \in L} x_{k,f,l} \frac{T}{T^{insp}_{l,f}}) \tag{D.4}$$

- Similarly, repair costs, maintenance costs and downtime caused by maintenance are related to specific failure modes in this case. Parameters $c\_repa_k$, $c\_main_k$ and $T\_main_k$ are rewritten as $c\_repa_{k,f}$, $c\_main_{k,f}$ and $T\_main_{k,f}$, respectively. It is worth mentioning that variable $mainTimes_k$, which is the ratio of the number of times of maintenance to the number of times of repair carried out for stage $k$ also breaks down to failure mode level and is rewritten as $mainTimes_{k,f}$. Accordingly, equations (57), (56), (58) and (59) are modified as (D.5),(D.6), (D.7) and (D.8) respectively.

$$mainRatio_{k,f} \geq y_{k,j}(\lambda^0_{k,j,f} - \lambda_{k,j,f})/\lambda_{k,j,f}, \quad \forall j \in J_k \tag{D.5}$$

$$repaCost = -T \sum_{\bar{s} \in \bar{S}} W(\bar{s}, \bar{s}) \pi_{\bar{s}} \sum_{(k,j,f) \in FAILURE_{\bar{s}}} c\_repa_{k,f} \tag{D.6}$$

$$mainCost = -T \sum_{\bar{s} \in \bar{S}} W(\bar{s}, \bar{s}) \pi_{\bar{s}} \sum_{(k,j,f) \in FAILURE_{\bar{s}}} mainRatio_{k,f} c\_main_{k,f} \tag{D.7}$$

$$mainTime = -T \sum_{\bar{s} \in \bar{S}^f} W(\bar{s}, \bar{s}) \pi_{\bar{s}} \sum_{(k,f) \in FAILURE_{\bar{s}}} mainRatio_{k,f} T\_main_{k,f} \tag{D.8}$$

## References


Aguilar, O., Kim, J.-K., Perry, S., and Smith, R. (2008). Availability and reliability considerations in the design and optimisation of flexible utility systems. *Chemical Engineering Science*, 63(14):3569–3584.

Alaswad, S. and Xiang, Y. (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. *Reliability Engineering & System Safety*, 157:54–63.

Bloch-Mercier, S. (2002). A preventive maintenance policy with sequential checking procedure for a markov deteriorating system. *European Journal of Operational Research*, 142(3):548–576.

Cheung, K.-Y., Hui, C.-W., Sakamoto, H., Hirata, K., and O'Young, L. (2004). Short-term site-wide maintenance scheduling. *Computers & chemical engineering*, 28(1):91–102.

Chiang, J.-H. and Yuan, J. (2001). Optimal maintenance policy for a markovian system under periodic inspection. *Reliability Engineering & System Safety*, 71(2):165–172.

Christer, A. (1999). Developments in delay time analysis for modelling plant maintenance. *Journal of the Operational Research Society*, pages 1120–1137.

Chryssaphinou, O., Limnios, N., and Malefaki, S. (2011). Multi-state reliability systems under discrete time semimarkovian hypothesis. *IEEE Transactions on reliability*, 60(1):80–87.

Ding, S.-H. and Kamaruddin, S. (2015). Maintenance policy optimizationliterature review and directions. *The International Journal of Advanced Manufacturing Technology*, 76(5-8):1263–1283.

Goel, H. D., Grievink, J., and Weijnen, M. P. (2003). Integrated optimal reliable design, production, and maintenance planning for multipurpose process plants. *Computers & Chemical Engineering*, 27(11):1543–1555.

Grossmann, I. E. and Trespalacios, F. (2013). Systematic modeling of discrete-continuous optimization models through generalized disjunctive programming. *AIChE Journal*, 59(9):3276–3295.

Jensen, H. A., Muñoz, A., Papadimitriou, C., and Millas, E. (2016). Model-reduction techniques for reliability-based design problems of complex structural systems. *Reliability Engineering & System Safety*, 149:204–217.

Kim, H. (2017). Optimal reliability design of a system with k-out-of-n subsystems considering redundancy strategies. *Reliability Engineering & System Safety*, 167:572 – 582.

Kuo, W. and Wan, R. (2007). Recent advances in optimal reliability allocation. In *Computational intelligence in reliability engineering*, pages 1–36. Springer.

Lee, H. and Cha, J. H. (2016). New stochastic models for preventive maintenance and maintenance optimization. *European Journal of Operational Research*, 255(1):80–90.

Lin, Z., Zheng, Z., Smith, R., and Yin, Q. (2012). Reliability issues in the design and optimization of process utility systems. *Theoretical Foundations of Chemical Engineering*, 46(6):747–754.

Moubray, J. (1997). *Reliability-centered maintenance*. New York : Industrial Press, 2nd ed edition. "RCM II"–Cover.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation.

Nguyen, D. and Bagajewicz, M. (2008). Optimization of preventive maintenance scheduling in processing plants. *Computer Aided Chemical Engineering*, 25:319–324.

Pistikopoulos, E. N., Vassiliadis, C. G., Arvela, J., and Papageorgiou, L. G. (2001). Interactions of maintenance and production planning for multipurpose process plants a system effectiveness approach. *Industrial & engineering chemistry research*, 40(14):3195–3207.

Raman, R. and Grossmann, I. E. (1991). Relation between milp modelling and logical inference for chemical process synthesis. *Computers & Chemical Engineering*, 15(2):73–84.

Sericola, B. (2013). *Markov Chains: Theory and Applications*. John Wiley & Sons.

Sharda, B. and Bury, S. J. (2008). A discrete event simulation model for reliability modeling of a chemical plant. In *Proceedings of the 40th Conference on Winter Simulation*, pages 1736–1740. Winter Simulation Conference.

Sharma, A., Yadava, G., and Deshmukh, S. (2011). A literature review and future perspectives on maintenance optimization. *Journal of Quality in Maintenance Engineering*, 17(1):5–25.

Shin, J. and Lee, J. H. (2016). Multi-time scale procurement planning considering multiple suppliers and uncertainty in supply and demand. *Computers & Chemical Engineering*, 91:114–126.

Shin, J., Lee, J. H., and Realff, M. J. (2017). Operational planning and optimal sizing of microgrid considering multi-scale wind uncertainty. *Applied energy*, 195:616–633.

Terrazas-Moreno, S., Grossmann, I. E., Wassick, J. M., and Bury, S. J. (2010). Optimal design of reliable integrated chemical production sites. *Computers & Chemical Engineering*, 34(12):1919–1936.

Thomaidis, T. and Pistikopoulos, E. (1994). Integration of flexibility, reliability and maintenance in process synthesis and design. *Computers & Chemical Engineering*, 18:S259–S263.

Thomaidis, T. V. and Pistikopoulos, E. (1995). Optimal design of flexible and reliable process systems. *IEEE transactions on reliability*, 44(2):243–250.

Vassilliadis, C. and Pistikopoulos, E. (1999). Chemical-process design and maintenance optimization under uncertainty: A simultaneous approach. In *Reliability and Maintainability Symposium, 1999. Proceedings. Annual*, pages 78–83. IEEE.

Weibull, W. et al. (1951). A statistical distribution function of wide applicability. *Journal of applied mechanics*, 18(3):293–297.

Ye, Y., Grossmann, I. E., and Pinto, J. M. (2017). Mixed-integer nonlinear programming models for optimal design of reliable chemical plants. *Computers & Chemical Engineering*.