Fault-tolerant checkpointing strategies with prediction

Anne Benoit **Michel Nicolis** Yves Robert Frédéric Vivien

ENS Lyon, LIP, ROMA

Fréjus, March 16, 2026

Table of Contents

Table of Contents

Problem statement

We have a platform with failures that occur following an exponential distribution of parameter $\lambda > 0$.

- We wish to know when to schedule checkpoints of duration C to save the work done

Problem statement

We have a platform with failures that occur following an exponential distribution of parameter $\lambda > 0$.

- We wish to know when to schedule checkpoints of duration C to save the work done
- We have a predictor that is characterised by a recall r, a precision p and a lead time ℓ

Previous work

This problem has been treated in a 2013 paper[1]. However there are key differences in the approach here:

[1] Checkpointing algorithms and fault prediction Guillaume Aupy, Yves Robert, Frédéric Vivien, Dounia Zaidouni Journal of Parallel and Distributed Computing, 2013, 74 (2), pp.2048-2064. ⟨10.1016/j.jpdc.2013.10.010⟩

[2] Checkpointing strategies to tolerate non-memoryless failures on HPC platforms Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien ACM Transactions on Parallel Computing, 2024, 11 (1), pp.1-26. ⟨10.1145/3624560⟩

Previous work

This problem has been treated in a 2013 paper[1]. However there are key differences in the approach here:

- The 2013 article sought first order approximations while we try to find an exact solution

[1] Checkpointing algorithms and fault prediction Guillaume Aupy, Yves Robert, Frédéric Vivien, Dounia Zaidouni Journal of Parallel and Distributed Computing, 2013, 74 (2), pp.2048-2064. ⟨10.1016/j.jpdc.2013.10.010⟩

[2] Checkpointing strategies to tolerate non-memoryless failures on HPC platforms Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien ACM Transactions on Parallel Computing, 2024, 11 (1), pp.1-26. ⟨10.1145/3624560⟩

Previous work

This problem has been treated in a 2013 paper[1]. However there are key differences in the approach here:

- The 2013 article sought first order approximations while we try to find an exact solution
- We take into account the lead time of the predictor

[1] Checkpointing algorithms and fault prediction Guillaume Aupy, Yves Robert, Frédéric Vivien, Dounia Zaidouni Journal of Parallel and Distributed Computing, 2013, 74 (2), pp.2048-2064. ⟨10.1016/j.jpdc.2013.10.010⟩

[2] Checkpointing strategies to tolerate non-memoryless failures on HPC platforms Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien ACM Transactions on Parallel Computing, 2024, 11 (1), pp.1-26. ⟨10.1145/3624560⟩

Previous work

This problem has been treated in a 2013 paper[1]. However there are key differences in the approach here:

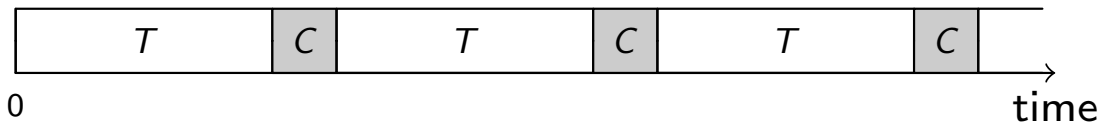
- The 2013 article sought first order approximations while we try to find an exact solution
- We take into account the lead time of the predictor
- Rather than minimizing the makespan: maximize the work completed before the first fault strikes[2]

[1] Checkpointing algorithms and fault prediction Guillaume Aupy, Yves Robert, Frédéric Vivien, Dounia Zaidouni Journal of Parallel and Distributed Computing, 2013, 74 (2), pp.2048-2064. ⟨10.1016/j.jpdc.2013.10.010⟩

[2] Checkpointing strategies to tolerate non-memoryless failures on HPC platforms Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien ACM Transactions on Parallel Computing, 2024, 11 (1), pp.1-26. ⟨10.1145/3624560⟩

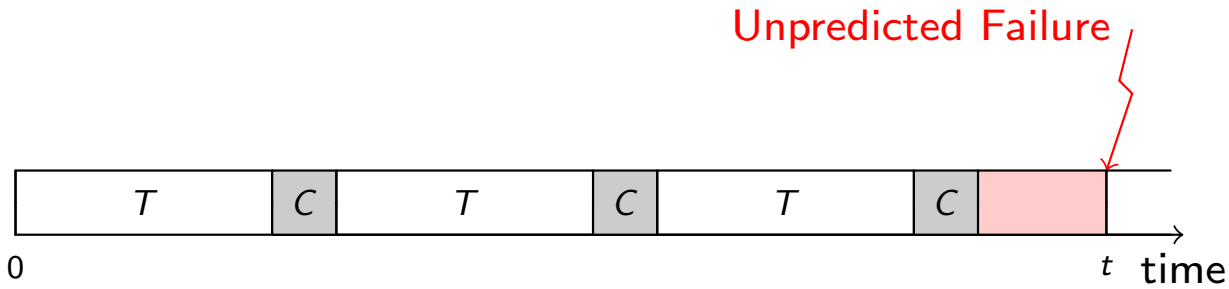
Case with no predictor

We checkpoint periodically every T units of work



Case with no predictor

When a fault strikes all of the work done since the last checkpoint is lost



The work done is thus T times the number of periods $T + C$ completed when the first fault strikes at time t

Work done in the case with no predictor

$$\left\lfloor \frac{t}{T+C} \right\rfloor T$$

Expectation of the work done in the case with no predictor

$$\mathbb{E}_{nopredictor} = \int_{t=0}^{t=+\infty} \left\lfloor \frac{t}{T+C} \right\rfloor T e^{-\lambda t} dt$$

Table of Contents

Strategy

A predictor with

- a precision $p = 1$
- a recall $r < 1$
- a lead time ℓ

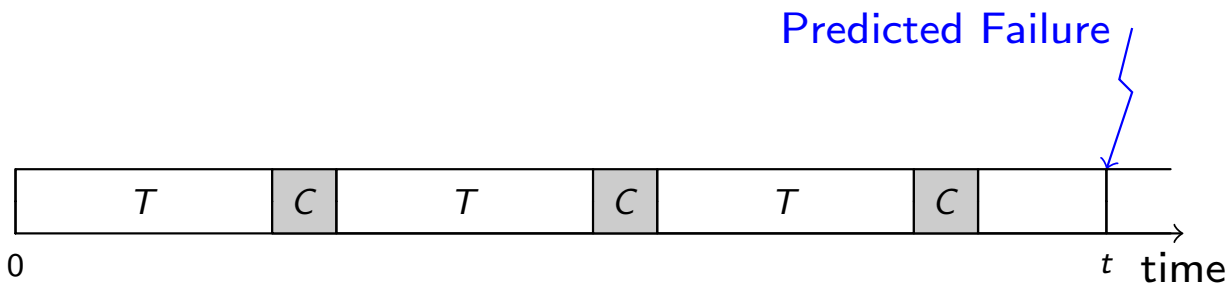
Strategy

A predictor with

- a precision $p = 1$
- a recall $r < 1$
- a lead time ℓ

When we have a prediction:

- We schedule a new checkpoint as late as possible



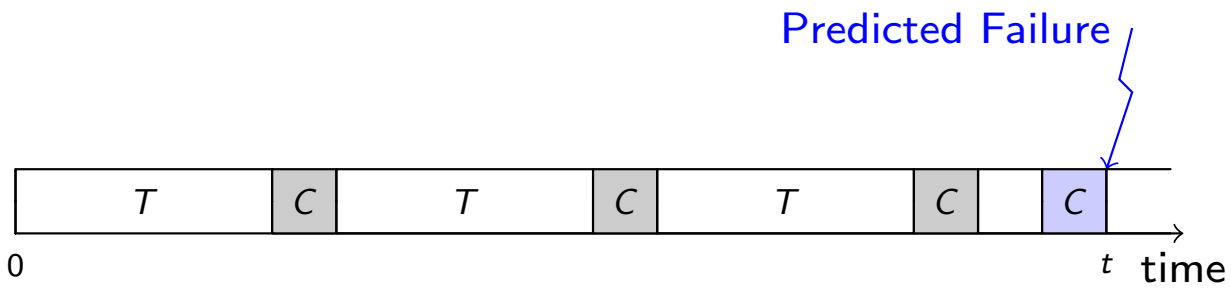
Strategy

A predictor with

- a precision $p = 1$
- a recall $r < 1$
- a lead time ℓ

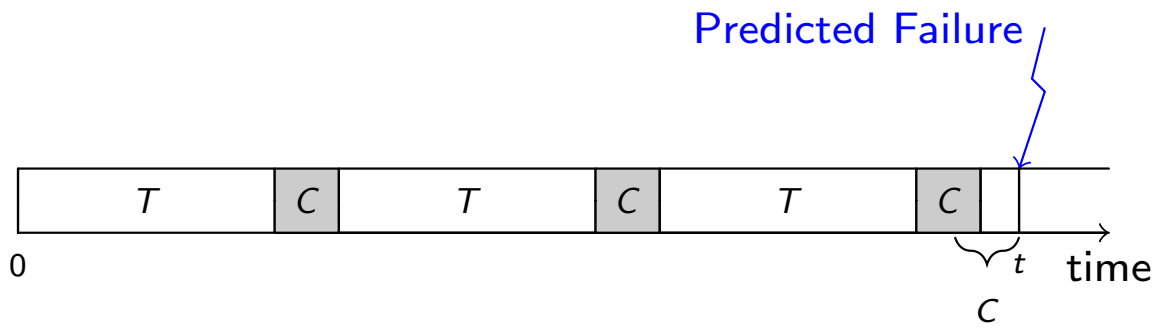
When we have a prediction:

- We schedule a new checkpoint as late as possible



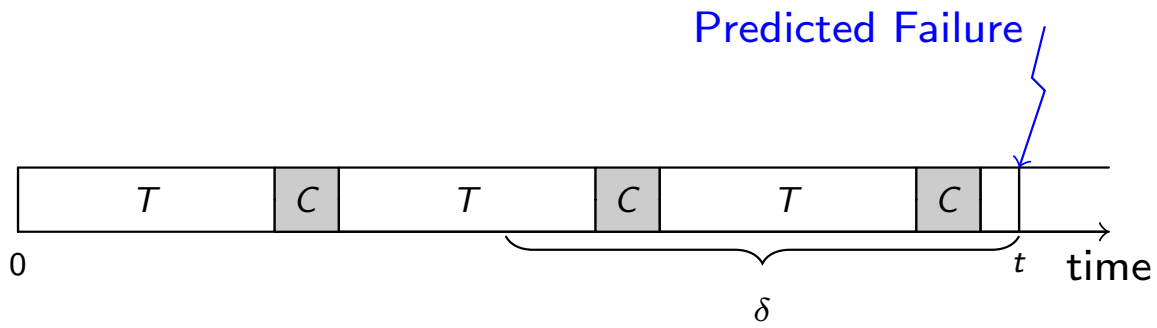
Strategy

- What happens if a fault is predicted to happen right before a periodic checkpoint?



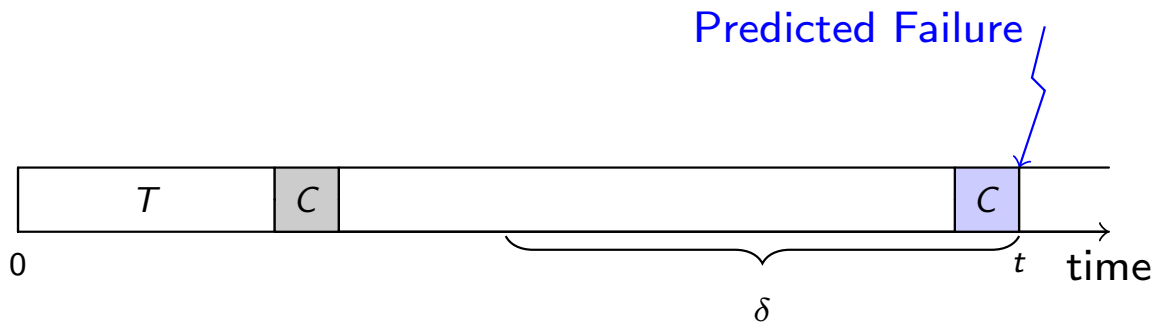
Strategy

- What happens if a fault is predicted to happen right before a periodic checkpoint?
- We introduce a δ such that $\delta \leq \ell$ and $\delta \geq 2C$ and cancel every periodic checkpoint that would happen within δ of the predicted fault



Strategy

- What happens if a fault is predicted to happen right before a periodic checkpoint?
- We introduce a δ such that $\delta \leq \ell$ and $\delta \geq 2C$ and cancel every periodic checkpoint that would happen within δ of the predicted fault



Expectation of the work done when the first fault strikes

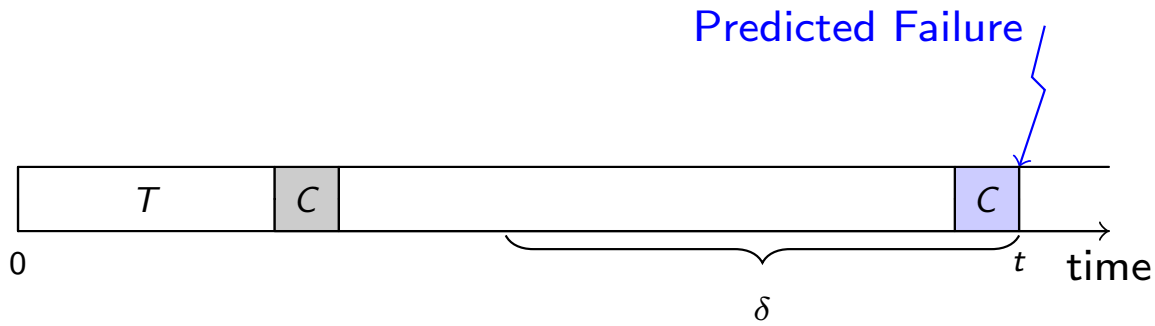
We separate the case where the first fault is predicted (probability r) and the case where the first fault is not predicted (probability $1 - r$)

Overall expectation of the work saved

$$\mathbb{E} = r\mathbb{E}_{pred} + (1 - r)\mathbb{E}_{unpred}$$

When the first fault is predicted

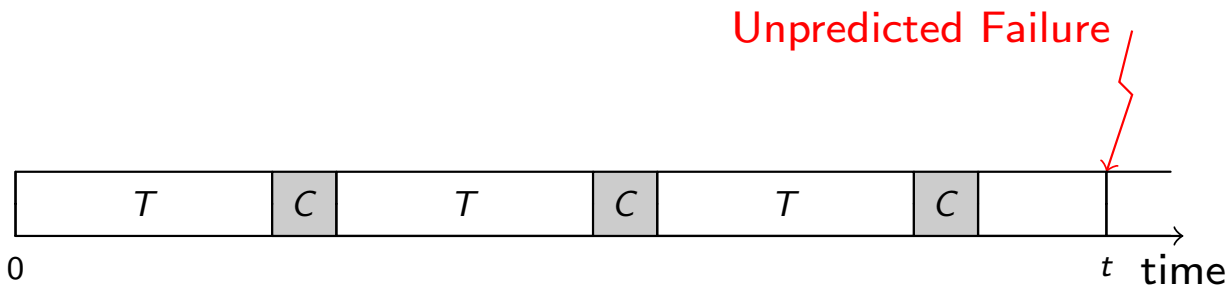
If the first fault is predicted then we perform a checkpoint every T units of work until δ before the fault strikes



Expectation of the work done when the first fault is predicted

$$\begin{aligned}\mathbb{E}_{pred} &= \int_{t=C}^{t=\delta} (t-C)\lambda e^{-\lambda t} dt + \int_{t=\delta}^{t=+\infty} \left(t - \left\lfloor \frac{t-\delta}{T+C} \right\rfloor C - C \right) \lambda e^{-\lambda t} dt \\ &= \frac{e^{-\lambda C}}{\lambda} - \frac{C e^{-\lambda \delta}}{e^{\lambda(T+C)} - 1}\end{aligned}$$

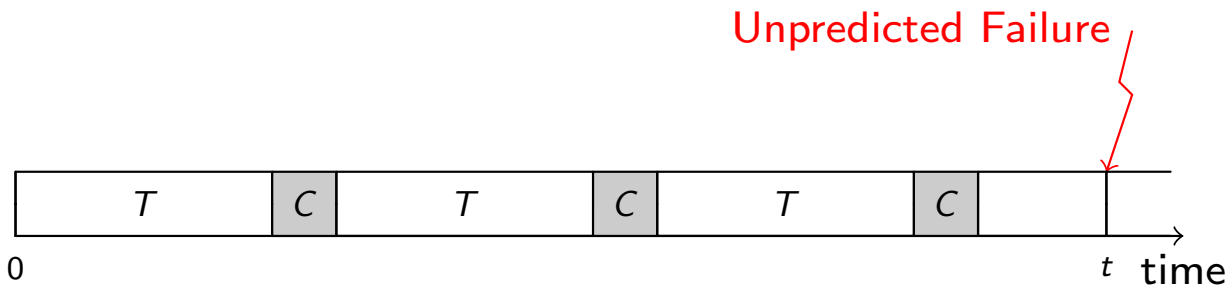
When the first fault is not predicted



Expectation of the work done in the case with no predictor

$$\mathbb{E}_{nopredictor} = \int_{t=0}^{t=+\infty} \left\lfloor \frac{t}{T+C} \right\rfloor T dt$$

When the first fault is not predicted



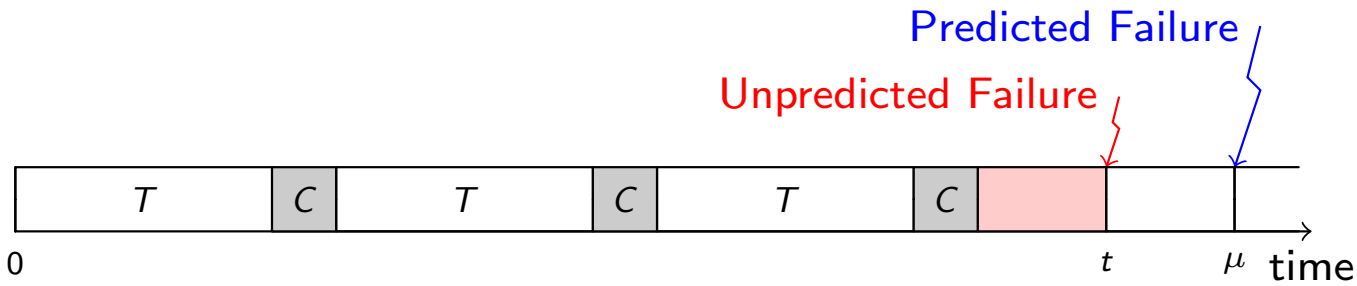
Expectation of the work done in the case with no predictor

$$\mathbb{E}_{\text{nopredictor}} = \int_{t=0}^{t=+\infty} \left\lfloor \frac{t}{T+C} \right\rfloor T \, \mathrm{d}t$$

This is false!

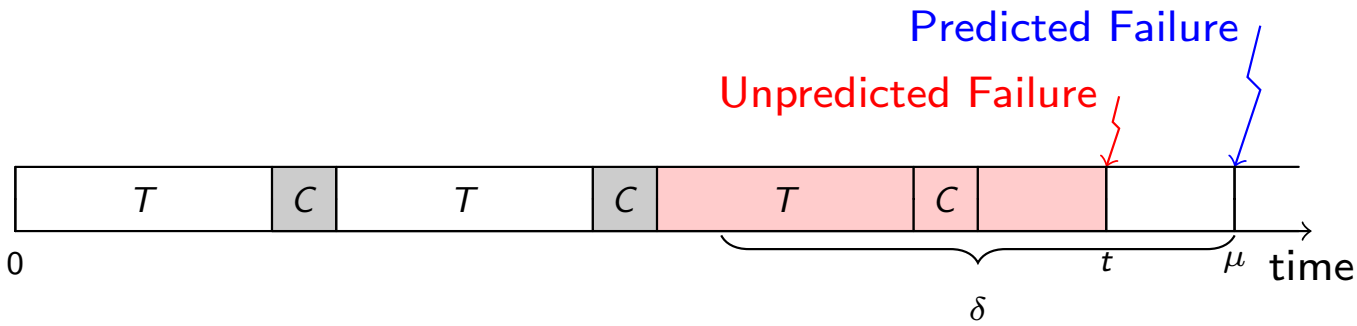
When the first fault is not predicted

Case where the first fault is not predicted but the second one is.



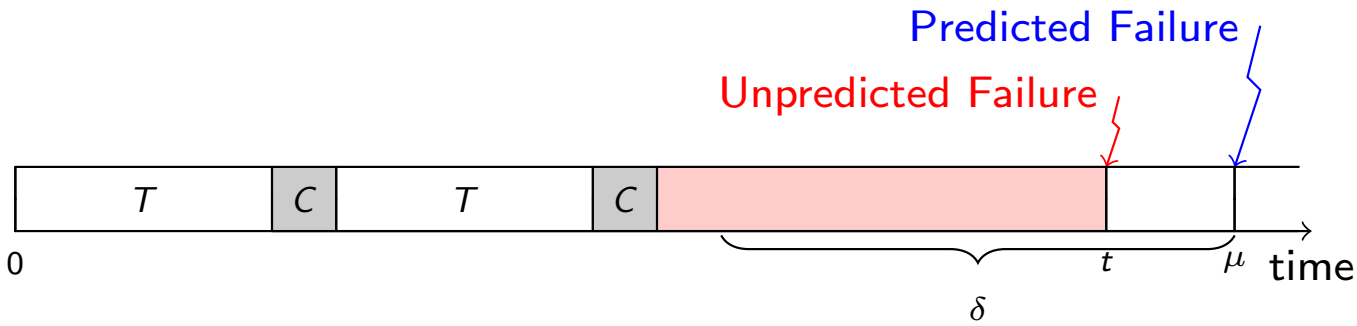
When the first fault is not predicted

Case where the first fault is not predicted but the second one is.



When the first fault is not predicted

Case where the first fault is not predicted but the second one is.



Because of our strategy we have saved less work in this case compared to the strategy that does not use a predictor.

When the first fault is not predicted

The probability that at least one fault is predicted to strike within δ of the first fault is given by a Poisson distribution of parameter $r\lambda$

Expectation of the work done when the first fault is not predicted

$$\mathbb{E}_{unpred} = e^{-r\lambda\delta}\mathbb{E}_{unpred1} + (1 - e^{-r\lambda\delta})\mathbb{E}_{unpred2}$$

When the first fault is not predicted

The first fault is not predicted and no fault is predicted in the next δ units of time

$$\mathbb{E}_{unpred1} = \int_{t=0}^{t=+\infty} \left(\left\lfloor \frac{t}{T+C} \right\rfloor T \right) \lambda e^{-\lambda t} dt$$

$$\mathbb{E}_{unpred1} = \frac{T}{e^{\lambda(T+C)} - 1}$$

When the first fault is not predicted

The first fault is not predicted and at least one fault is predicted in the next δ units of time

$$\mathbb{E}_{unpred2} = \int_{t=0}^{t=+\infty} \int_{\mu=t}^{\mu=t+\delta} \frac{\lambda r e^{-\lambda r \mu}}{1 - e^{-\lambda r \delta}} \left\lfloor \frac{max(0, \mu - \delta)}{T + C} \right\rfloor T d\mu \lambda e^{-\lambda t} dt$$

$$\mathbb{E}_{unpred2} = \frac{T r e^{-\lambda(r+1)\delta}(e^{\delta\lambda} - 1)}{(1 - e^{-r\lambda\delta})(r+1)(e^{\lambda(r+1)(T+C)} - 1)}$$

When the first fault is not predicted

The first fault is not predicted

$$\mathbb{E}_{unpred} = e^{-r\lambda\delta}\mathbb{E}_{unpred1} + (1 - e^{-r\lambda\delta})\mathbb{E}_{unpred2}$$

$$\mathbb{E}_{unpred} = e^{-r\lambda\delta}\frac{T}{e^{\lambda(T+C)} - 1} + \frac{Tre^{-\lambda(r+1)\delta}(e^{\delta\lambda} - 1)}{(r+1)(e^{\lambda(r+1)(T+C)} - 1)}$$

Expectation of the work done when the first fault strikes

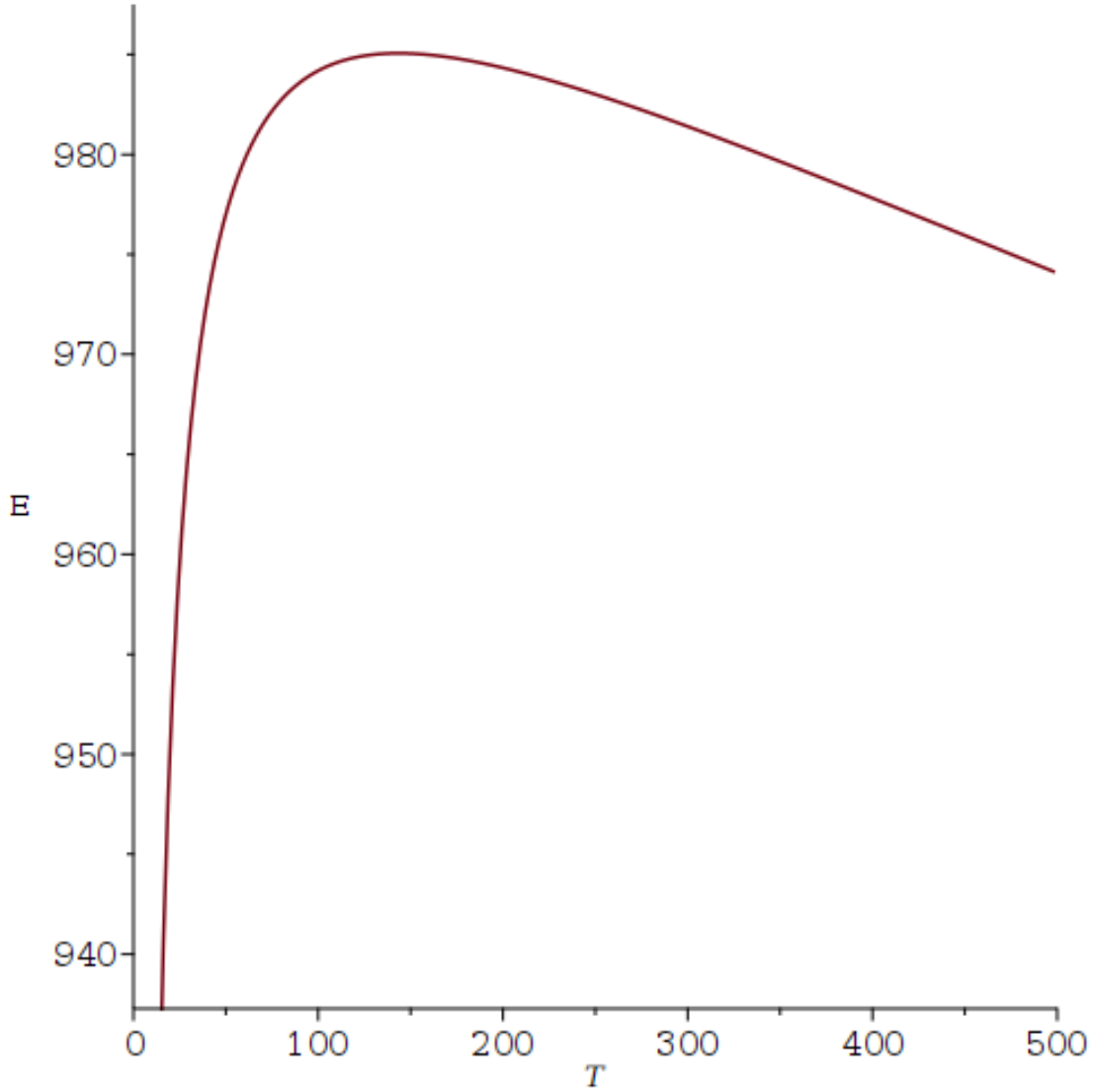
Expectation of the work done when the first fault strikes

$$\mathbb{E} = r\mathbb{E}_{pred} + (1-r)\mathbb{E}_{unpred}$$

$$\mathbb{E} = r\left(\frac{e^{-\lambda C}}{\lambda} - \frac{Ce^{-\lambda\delta}}{e^{\lambda(T+C)}-1}\right) + (1-r)\left(e^{-r\lambda\delta}\frac{T}{e^{\lambda(T+C)}-1} + \frac{Tre^{-\lambda(r+1)\delta}(e^{\delta\lambda}-1)}{(r+1)(e^{\lambda(r+1)(T+C)}-1)}\right)$$

Table of Contents

Results and future work



Optimal T?

Optimal δ?

$p < 1$?

Thank you for your attention.