# Combining Process Replication and Checkpointing for Resilience

Henri Casanova[1], Yves Robert[2,3,4], Frédéric Vivien[5,2], and Dounia Zaidouni[5,2]

1. University of Hawai'i
2. Ecole Normale Supérieure de Lyon
3. Institut Universitaire de France
4. University of Tennessee Knoxville
5. INRIA

June 14, 2012

# How to address fault-tolerance at exascale?

Most classical approach: rollback-recovery

- What is the most appropriate protocol?
  (cf. yesterday's talk by Amina Guermouche)
- How efficient will checkpointing protocols be?
- Can some external mechanisms improve efficiency and resilience of checkpointing protocols?

# How to address fault-tolerance at exascale?

Most classical approach: rollback-recovery

- What is the most appropriate protocol?
  (cf. yesterday's talk by Amina Guermouche)
- How efficient will checkpointing protocols be?
- Can some external mechanisms improve efficiency and resilience of checkpointing protocols?

Alternative approach: replication

- Systematic replication: efficiency $< 50\%$
- Can replication+checkpointing be more efficient than checkpointing alone?
- Claim by Ferreira et al. [Supercomputing 2011]: yes

Our aim: revisit their study

# Outline

1. Process replication
2. Combining process replication and checkpointing
3. Conclusion

# Outline

1. Process replication
   - Model
   - Analogy with birthday problem (when $g = 2$)
   - Computing the MTTI
   - Numerical evaluation
2. Combining process replication and checkpointing
   - Impact of checkpointing period
   - Evaluating replication
3. Conclusion

# Outline

# Model by Ferreira et al. [Supercomputing 2011]

- A parallel application comprising $n$ (sequential) processes
- Each process replicated $g \geq 2$ times $\rightarrow$ *replica-group*
- A processing element executes a single replica
  Two replicas, even from two different application processes, cannot run on the same PE
- When a replica is hit by a failure, it is not restarted
  Underlying assumption: the whole application runs at the speed of the lowest replica
- The application fails when all replicas in one replica-group have been hit by failures
- Failures of different PEs are not correlated
- Study for $g = 2$ by Ferreira et al., SC'2011

# Question: what is the value of the MNFTI?

What is the mean number of processing element failures needed to interrupt the application?

In other words: What is the mean number of processing element failures needed to kill all replicas in (at least) one replica-group?

# Outline

# The birthday problem

## Classical formulation

What is the probability, in a set of $m$ people, that two of them have same birthday ?

## Relevant formulation

What is the average number of people required to find a pair with same birthday?

$$F(n) = 1 + \sum_{k=1}^{n} \frac{n!}{(n-k)! \cdot n^k}$$

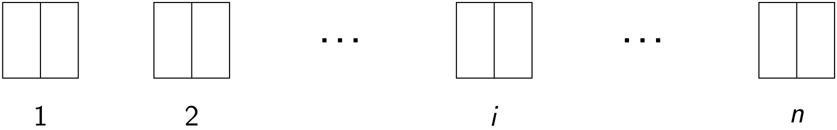
## The analogy

Two people with same birthday

$$\equiv$$

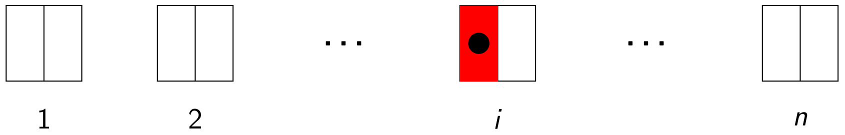
Two failures hitting same replica-group

# Differences with birthday problem



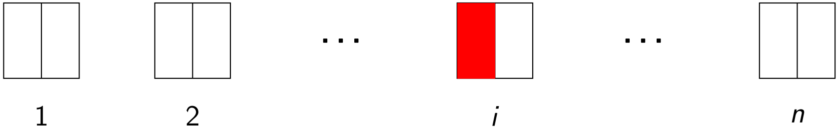

- $n$ processes; each replicated twice
- Uniform distribution of failures

# Differences with birthday problem



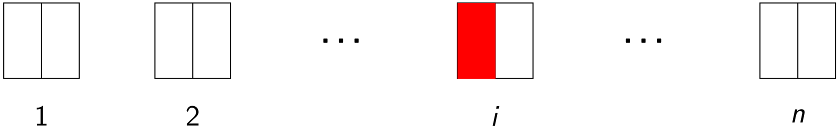

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)

# Differences with birthday problem



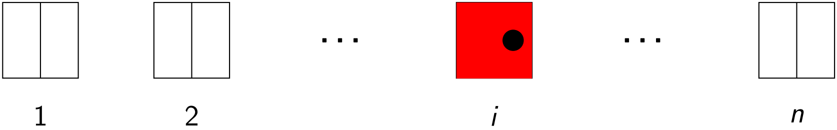

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure: can the failed PE be hit?

# Differences with birthday problem



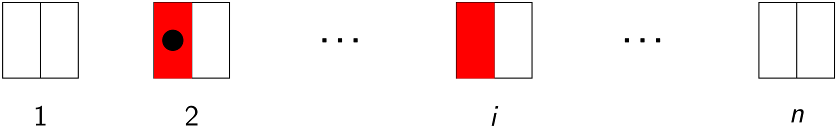

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure cannot hit failed PE
  - Failure uniformly distributed over $2n - 1$ PEs

# Differences with birthday problem



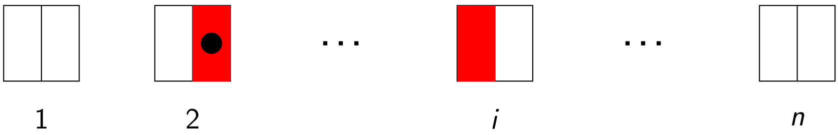

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure cannot hit failed PE
  - Failure uniformly distributed over $2n-1$ PEs
  - Probability that replica-group $i$ is hit by failure: $1/(2n-1)$

# Differences with birthday problem



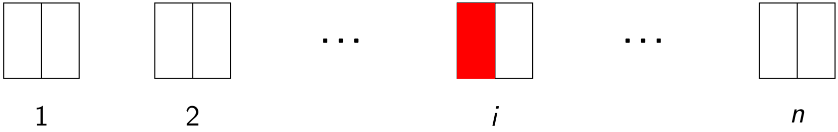

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure cannot hit failed PE
  - Failure uniformly distributed over $2n-1$ PEs
  - Probability that replica-group $i$ is hit by failure: $1/(2n-1)$
  - Probability that replica-group $\neq i$ is hit by failure: $2/(2n-1)$

# Differences with birthday problem



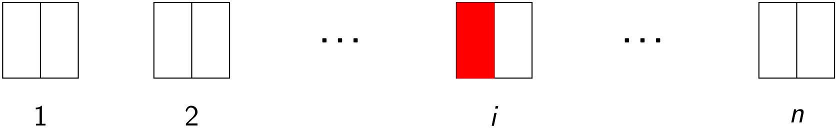

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure cannot hit failed PE
  - Failure uniformly distributed over $2n - 1$ PEs
  - Probability that replica-group $i$ is hit by failure: $1/(2n - 1)$
  - Probability that replica-group $\neq i$ is hit by failure: $2/(2n - 1)$

# Differences with birthday problem



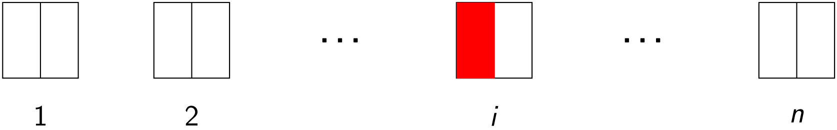

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure cannot hit failed PE
  - Failure uniformly distributed over $2n-1$ PEs
  - Probability that replica-group $i$ is hit by failure: $1/(2n-1)$
  - Probability that replica-group $\neq i$ is hit by failure: $2/(2n-1)$
  - Failure not uniformly distributed over replica-groups: this is not the birthday problem

# Differences with birthday problem



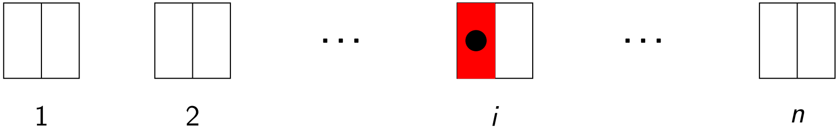

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure can hit failed PE

# Differences with birthday problem



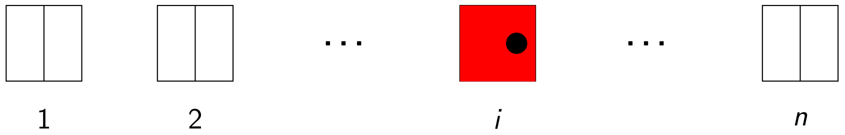

- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure can hit failed PE
  - Suppose the failure hit replica-group $i$

# Differences with birthday problem




- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure can hit failed PE
  - Suppose the failure hit replica-group $i$
  - If the failure hit the failed PE: application survives

# Differences with birthday problem




- $n$ processes; each replicated twice
- Uniform distribution of failures
- First failure: each replica-group has probability $1/n$ to be hit
- Nothing is restarted (neither on failed PE nor elsewhere)
- Second failure can hit failed PE
  - Suppose the failure hit replica-group $i$
  - If the failure hit the failed PE: application survives
  - If the failure hit the running PE: application killed
  - Not all failures hitting the same replica-group are equal: this is not the birthday problem

# Computing $MNFTI^{\mathrm{rp}}$ (1/4)

- Hypothesis: failures can only hit running PEs
- Each application process has 2 replicas: $g = 2$

- $n_f$: number of replica-groups already hit by failures
  - $n_f$ PEs have failed
  - $2n - n_f$ PEs still running

# Computing $MNFTI^{\mathrm{rp}}$ (2/4)

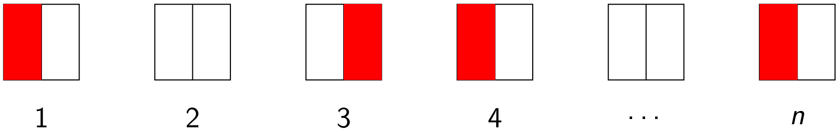
Case $n_f = n$

Next PE failure induces application failure

$$\mathbb{E}(NFTI^{\mathrm{rp}}|n) = 1$$

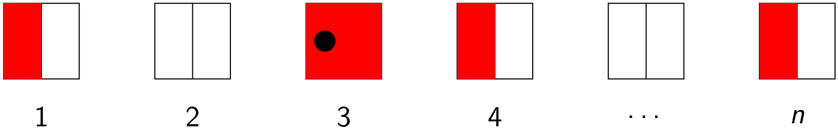
# Computing $MNFTI^{\mathrm{rp}}$ (3/4)

## General case



# Computing $MNFTI^{\text{rp}}$ (3/4)

## General case



Failure hit one of the $n_f$ already hit replica-groups

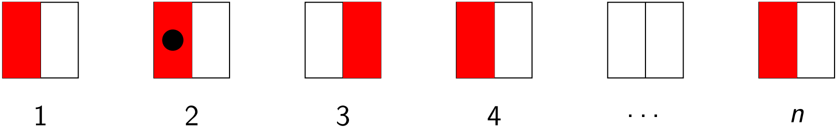
Probability: $\frac{n_f}{2n - n_f}$

Average number of failures needed for the application to fail:

$$1$$

# Computing $MNFTI^{\mathrm{rp}}$ (3/4)

## General case




Failure hit one replica-group with two running PEs

Probability: $\dfrac{2(n - n_f)}{2n - n_f}$

Average number of failures needed for the application to fail:

$$1 + \mathbb{E}(NFTI^{\mathrm{rp}} | n_f + 1)$$

# Computing $MNFTI^{\mathrm{rp}}$ (4/4)

**Theorem**

*If the failure inter-arrival times on the different PEs are i.i.d. then using process replication with* $g = 2$, $MNFTI^{\mathrm{rp}} = \mathbb{E}(NFTI^{\mathrm{rp}}|0)$ *where*

$$\mathbb{E}(NFTI^{\mathrm{rp}}|n_f) = \begin{cases} 1 & \text{if } n_f = n, \\ 1 + \frac{2n-2n_f}{2n-n_f}\mathbb{E}(NFTI^{\mathrm{rp}}|n_f+1) & \text{otherwise.} \end{cases}$$

# Computing $MNFTI^{\mathrm{rp}}$ (4/4)

**Theorem**

*If the failure inter-arrival times on the different PEs are i.i.d. then using process replication with* $g = 2$, $MNFTI^{\mathrm{rp}} = \mathbb{E}(NFTI^{\mathrm{rp}}|0)$ *where*

$$\mathbb{E}(NFTI^{\mathrm{rp}}|n_f) = \begin{cases} 1 & \text{if } n_f = n, \\ 1 + \frac{2n-2n_f}{2n-n_f}\mathbb{E}(NFTI^{\mathrm{rp}}|n_f+1) & \text{otherwise.} \end{cases}$$

**Theorem**

*If the failure inter-arrival times on the different PEs are i.i.d. and independent from the PE failure history, then*

$$MNFTI^{\mathrm{ah}} = 1 + MNFTI^{\mathrm{rp}}$$

# Generalization to any value of $g$

**Theorem**

*If the failure inter-arrival times on the different PEs are i.i.d.*

$$MNFTI^{\mathrm{rp}} = \mathbb{E}\left( NFTI^{\mathrm{rp}} | \underbrace{0, \ldots, 0}_{g-1 \text{ zeros}} \right) \text{ where:}$$

$$\mathbb{E}\left(NFTI^{\mathrm{rp}} | n_f^{(1)}, \ldots, n_f^{(g-1)}\right) = 1$$
$$+ \frac{g \cdot \left(n - \sum_{i=1}^{g-1} n_f^{(i)}\right)}{g \cdot n - \sum_{i=1}^{g-1} i \cdot n_f^{(i)}} \cdot \mathbb{E}\left(NFTI^{\mathrm{rp}} | n_f^{(1)}, n_f^{(2)}, \ldots, n_f^{(g-1)}\right)$$
$$+ \sum_{i=1}^{g-2} \frac{(g-i) \cdot n_f^{(i)}}{g \cdot n - \sum_{i=1}^{g-1} i \cdot n_f^{(i)}}$$
$$\cdot \mathbb{E}\left(NFTI^{\mathrm{rp}} | n_f^{(1)}, \ldots, n_f^{(i-1)}, n_f^{(i)} - 1,\ n_f^{(i+1)} + 1, n_f^{(i+2)}, \ldots, n_f^{(g-1)}\right)$$

# Outline

1. Process replication
   - Model
   - Analogy with birthday problem (when $g = 2$)
   - Computing the MTTI
   - Numerical evaluation
2. Combining process replication and checkpointing
   - Impact of checkpointing period
   - Evaluating replication
3. Conclusion

# From MNFTI to MTTI

$$MTTI = systemMTBF(g \cdot n) \times MNFTI^{\mathrm{ah}}(n)$$

True for exponential distribution of failures

What about other distributions?

# MTTI for any failure distribution

**$R(t)$ probability that application still running at time $t$**

- All replica-groups have at least one replica running
- Exponential: $R(t) = \left(1 - \left(1 - e^{-\lambda t}\right)^g\right)^n$
- Weibull: $R(t) = \left(1 - \left(1 - e^{-\left(\frac{t}{\lambda}\right)^k}\right)^g\right)^n$

*$MTTI$*

- $MTTI = \int_0^{+\infty} R(t)\mathrm{d}t$ $\quad\rightarrow$ closed-form formulas

# Outline

1. Process replication
   - Model
   - Analogy with birthday problem (when $g = 2$)
   - Computing the MTTI
   - Numerical evaluation
2. Combining process replication and checkpointing
   - Impact of checkpointing period
   - Evaluating replication
3. Conclusion

# Numerical evaluation of the MNFTI

| Number of processes $n$ | $2^0$ | $2^1$ | $2^2$ | $2^3$ | $2^4$ | $2^5$ | $2^6$ |
|---|---|---|---|---|---|---|---|
| Ferreira et al. | 2 | 2.5 | 3.22 | 4.25 | 5.7 | 7.77 | 10.7 |
| This work | 3 | 3.67 | 4.66 | 6.09 | 8.15 | 11.1 | 15.2 |
| % Relative Difference | -33 | -32 | -31 | -30 | -30 | -30 | -30 |

| Number of processes $n$ | $2^7$ | $2^8$ | $2^9$ | $2^{10}$ | $2^{11}$ | $2^{12}$ | $2^{13}$ |
|---|---|---|---|---|---|---|---|
| Ferreira et al. | 14.9 | 20.7 | 29 | 40.8 | 57.4 | 80.9 | 114 |
| This work | 21.1 | 29.4 | 41.1 | 57.7 | 81.2 | 114 | 161 |
| % Relative Difference | –30 | -29 | -29 | -29 | -29 | -29 | -29 |

| Number of processes $n$ | $2^{14}$ | $2^{15}$ | $2^{16}$ | $2^{17}$ | $2^{18}$ | $2^{19}$ | $2^{20}$ |
|---|---|---|---|---|---|---|---|
| Ferreira et al. | 161 | 228 | 322 | 454 | 642 | 908 | 1284 |
| This work | 228 | 322 | 455 | 643 | 908 | 1284 | 1816 |
| % Relative Difference | -29 | -29 | -29 | -29 | -29 | -29 | -29 |

# Outline

# Outline

# The question

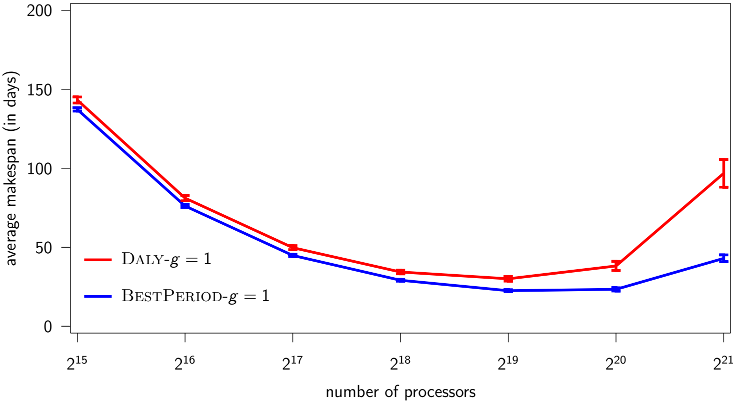
Ferreira et al. use Daly's checkpointing period

(without and with replication)

Does this matter?

# Without replication

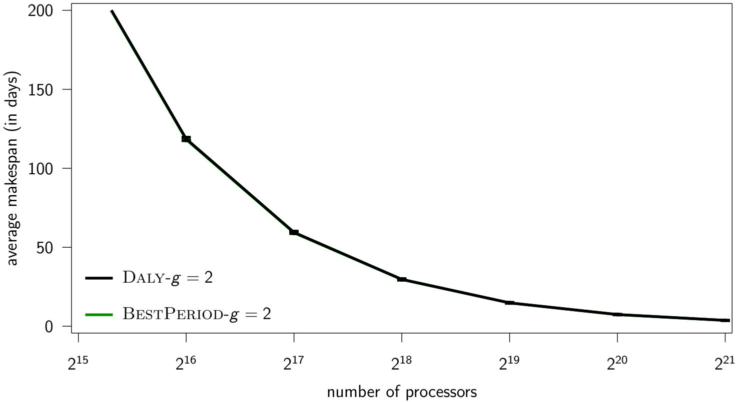
Weibull distribution with $k = 0.7$, PE MTBF of 125 years




The checkpointing period can have a significant impact

# With replication

Weibull distribution with $k = 0.7$, PE MTBF of 125 years




Daly's period appears to be an excellent choice

# Checkpoints are almost useless with replication

Weibull distribution

| | # of application failures | | % of PE failures | |
|---|---|---|---|---|
| # of processes | $k = 0.7$ | $k = 0.5$ | $k = 0.7$ | $k = 0.5$ |
| $2^{14}$ | 1.95 | 4.94 | 0.35 | 0.39 |
| $2^{15}$ | 1.44 | 3.77 | 0.25 | 0.28 |
| $2^{16}$ | 0.88 | 2.61 | 0.15 | 0.19 |
| $2^{17}$ | 0.45 | 1.67 | 0.075 | 0.12 |
| $2^{18}$ | 0.20 | 1.11 | 0.034 | 0.076 |
| $2^{19}$ | 0.13 | 0.72 | 0.022 | 0.049 |
| $2^{20}$ | 0.083 | 0.33 | 0.014 | 0.023 |

- Applications rarely rollback
- Daly's approximation is good enough

# Outline

# Methodology

## Ferreira et al.

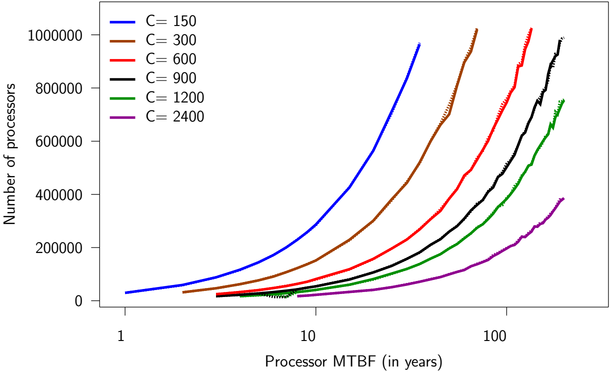
- Compare checkpointing without and with replication using Daly's period
- Problem: when $g = 1$ Daly's period may be suboptimal
- Conclusion: shows when replication is beneficial to Daly's periodic checkpointing

## Our approach

- Compare checkpointing without and with replication using best period
- Conclusion: shows when replication is beneficial to periodic checkpointing

# Exponential distribution

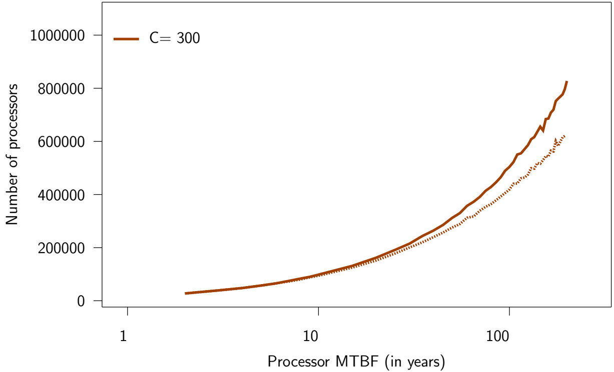
Dashed line: Ferreira et al.        Solid line: this work




- No difference between both approaches
- Replication beneficial if MTBF is low enough, checkpoints are large enough, the number of PEs is large enough

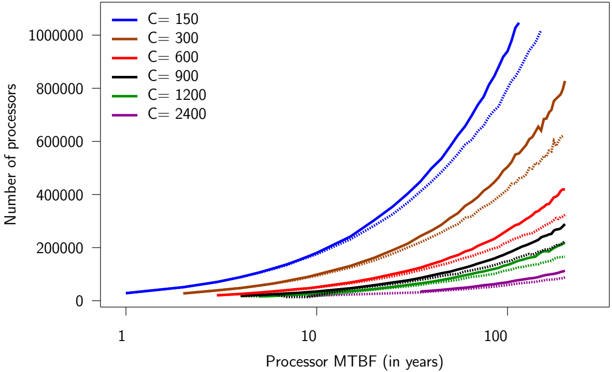
# Weibull distribution with $k = 0.7$

Dashed line: Ferreira et al. Solid line: this work



# Weibull distribution with $k = 0.7$

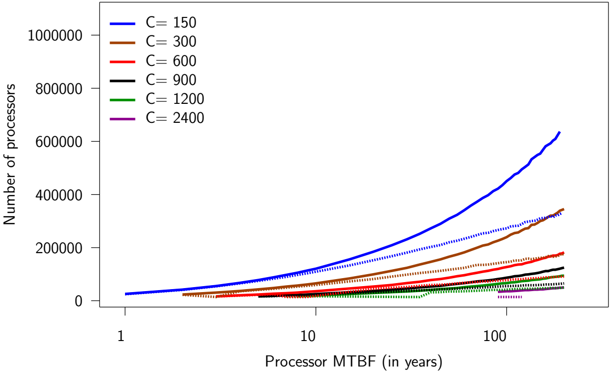
Dashed line: Ferreira et al. Solid line: this work




- Significant difference between both approaches
- Other conclusions are still valid

# Weibull distribution with $k = 0.5$

Dashed line: Ferreira et al. Solid line: this work




- Significant difference between both approaches
- Other conclusions are still valid

# Outline

# Conclusion

- Theoretical study by Ferreira et al. was flawed
- In practice, the theoretical flaw has no impact
- Simulation study by Ferreira et al. was flawed
- The flaw favored replication
- Depending of the failure distribution, replication can be quite less interesting than predicted Ferreira et al.
- Main flaws of this study:
  - Non correlated failures
  - Coordinated checkpointing