Replication for resilience @ exascale

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

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

http://graal.ens-lyon.fr/~yrobert/slides/2012-01-13-Knoxville.pdf
LAWN 262

Knoxville, January 13, 2012

Exascale platforms

- Massively parallel
  - 10^5 or 10^6 nodes
  - Each node equipped with 10^4 or 10^3 cores

- Failure-prone
  MTTF of one node $= 1$ year
  $\Rightarrow$ MTTF of platform with 10^6 nodes $= 30$sec

Exascale platforms

- Massively parallel
  - 10^5 or 10^6 nodes
  - Each node equipped with 10^4 or 10^3 cores

- Failure-prone
  MTTF of one node $= 1$ year
  $\Rightarrow$ MTTF of platform with 10^6 nodes $= 30$sec

  Exascale
  $\neq$ Petascale $\times 1000$

Outline

Outline

1 Best resource usage

2 Group replication

3 Process replication

4 Conclusion

Checkpointing

- C: checkpoint save time (in minutes)
- R: checkpoint recovery time (in minutes)
- D: down/reboot time (in minutes)
- μ: MTTF, mean time to failure
  (e.g., $1/\lambda$ if failures are exponentially distributed)
- N: total number of nodes

Failures (1/2)

- Exponential: density $p(t) = \lambda e^{-\lambda t}$
  $\Rightarrow \mu = 1/\lambda$

- Weibull: density $p(t) = (k/\lambda)(t/\lambda)^{k-1} e^{-(t/\lambda)^k}$
  $\Rightarrow \mu = \lambda \Gamma(1 + 1/k)$
  $\Rightarrow$ take $k = 0.5$ or $k = 0.7$ (values from literature)

- Values of MTTF
  - $\mu = 1$ year for ASCI Q machine
  - $\mu = 10\text{-}100$ years for Jaguar

Failures (2/2)

- If a job uses 2^j processors, what is the expected interval time between failures?

- μ_j mean of the minimum of 2^j i.i.d. variables
- Exponential with scale parameter λ:

$$\mu_j = 1/(\lambda 2^j) = \mu/2^j$$

- Weibull with scale parameter λ and shape parameter a:

$$\mu_j = \lambda \Gamma(1 + 1/(a2^j))$$

When to checkpoint?

Sequential jobs

- Exponential: best strategy periodic; Young, Daly, optimal, . . .
- Weibull: many heuristics

Parallel jobs

- No optimality result known
- Periodic heuristics available for Exponential and Weibull

Best resource usage

Context

- Large divisible job
- Large number N of identical nodes (same processing speed, same failure distribution)

Without failures

- More processors $\Rightarrow$ smaller makespan (so use N processors)

With failures

- More processors = more parallelism ☺
- More processors = more failures ☹
- How many processors ($\leq N$) to minimize expected makespan?

Scenarios

Jobs

- Can execute on any number $q \leq N$ processors
- Tightly coupled: all q processors operate synchronously
- Perfectly parallel jobs: $\mathcal{W}(q) = \mathcal{W}/q$.
- Generic parallel jobs: $\mathcal{W}(q) = \mathcal{W}/q + \gamma\mathcal{W}$
- Numerical kernels: $\mathcal{W}(q) = \mathcal{W}/q + \gamma\mathcal{W}^{2/3}/\sqrt{q}$

Checkpoint overhead

- Proportional overhead: $C(q) = R(q) = \alpha V/q = C/q$
  (bandwidth of processor network card/link is I/O bottleneck)
- Constant overhead: $C(q) = R(q) = \alpha V = C$
  (bandwidth to/from resilient storage system is I/O bottleneck)

Scenarios

Jobs

Theorem: for all scenarios, there is
an optimal (finite) number of processors

- Generic parallel jobs: $\mathcal{W}(q) = \mathcal{W}/q + \gamma\mathcal{W}$
- Numerical kernels: $\mathcal{W}(q) = \mathcal{W}/q + \gamma\mathcal{W}^{2/3}/\sqrt{q}$

Checkpoint overhead

- Proportional overhead: $C(q) = R(q) = \alpha V/q = C/q$
  (bandwidth of processor network card/link is I/O bottleneck)
- Constant overhead: $C(q) = R(q) = \alpha V = C$
  (bandwidth to/from resilient storage system is I/O bottleneck)

Scenarios

Jobs

Theorem: for all scenarios, there is
an optimal (finite) number of processors

- Generic parallel jobs: $W(q) = W/q + \gamma W$

- Numerical kernels: $W(q) = W/q + \gamma W^{2/3}/\sqrt{q}$

Use **replication** to benefit from all resources!

- Proportional overhead: $C(q) = R(q) = \alpha V/q = C/q$
  (bandwidth of processor network card/link is I/O bottleneck)

- Constant overhead: $C(q) = R(q) = \alpha V = C$
  (bandwidth to/from resilient storage system is I/O bottleneck)

Simulation setting

- $C = R = 10mn$ & $D = 1mn$ (Local SDD, scenario "2012")
- $N = 2^{20}$ nodes
- $\mathcal{W} = 1,000$ years (9 hours on whole failure-free platform)
- MTTF $\mu = 10$ years

Exponential distribution (MTTF $\mu = 10$ years)
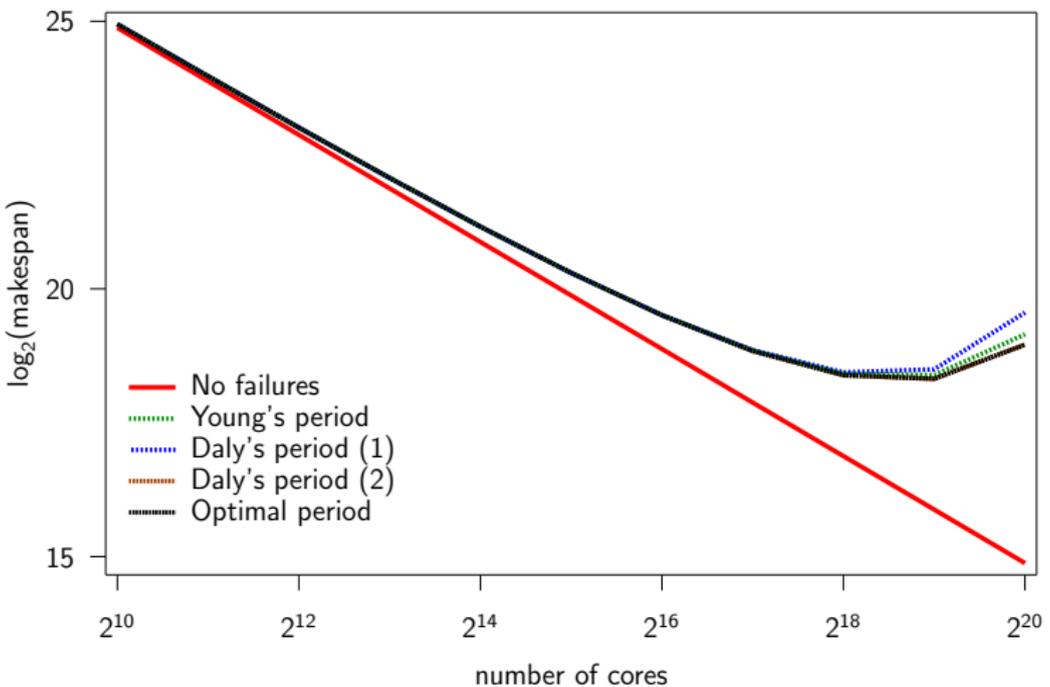
Best makespan with 2^{19} nodes instead of 2^{20} nodes

Outline

1 Best resource usage

2 Group replication

3 Process replication

4 Conclusion

Hypotheses

Parallelization

- Perfectly parallel jobs: $\mathcal{W}(q) = \mathcal{W}/q$
  (even more striking with Amdhal jobs or numerical kernels)

Replication

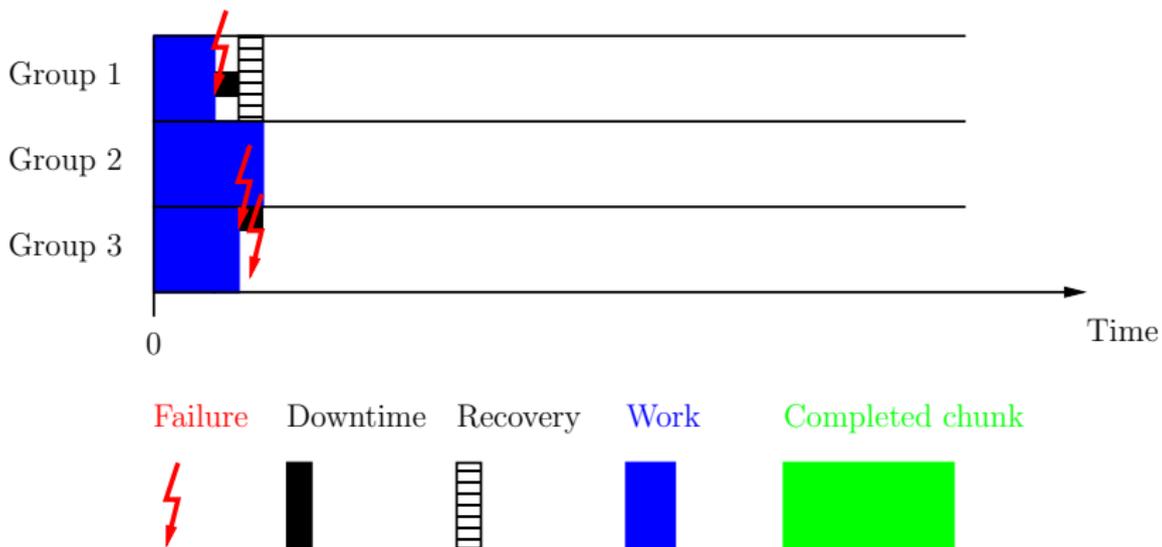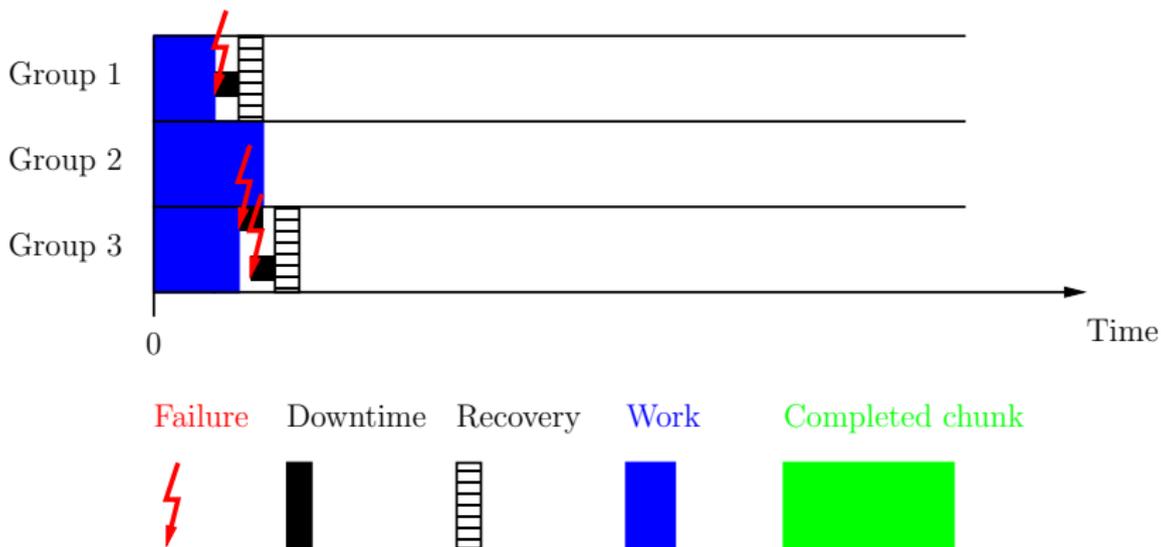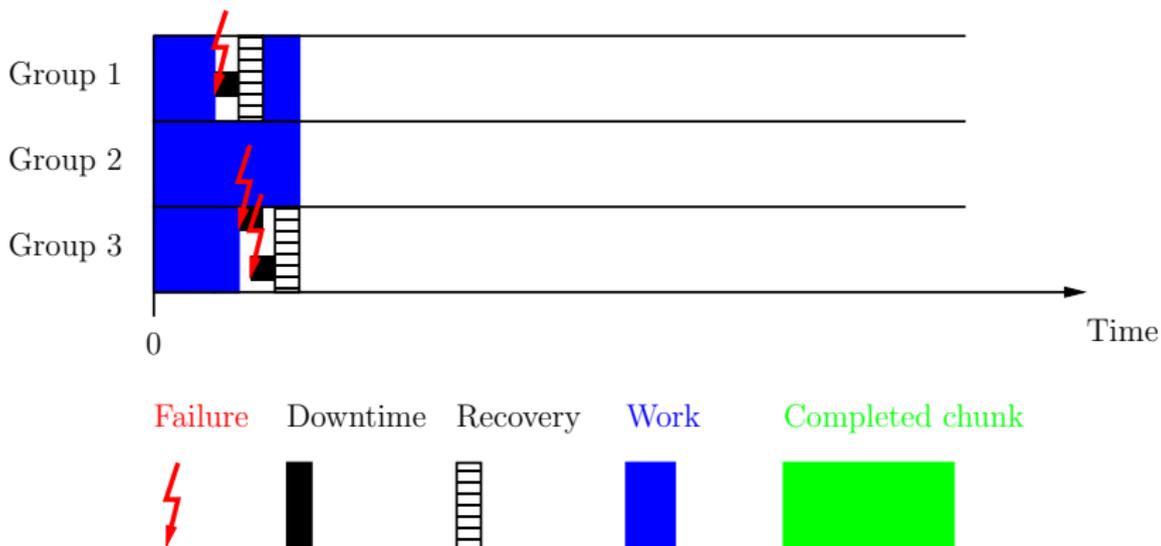- g groups of p processors ($g \times p \leq N$)

Group execution of a chunk
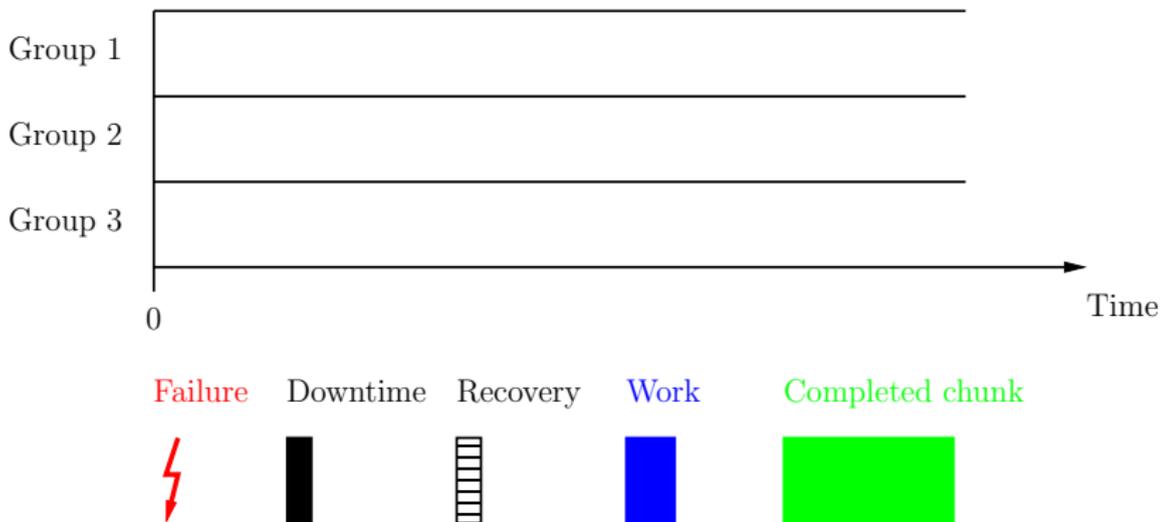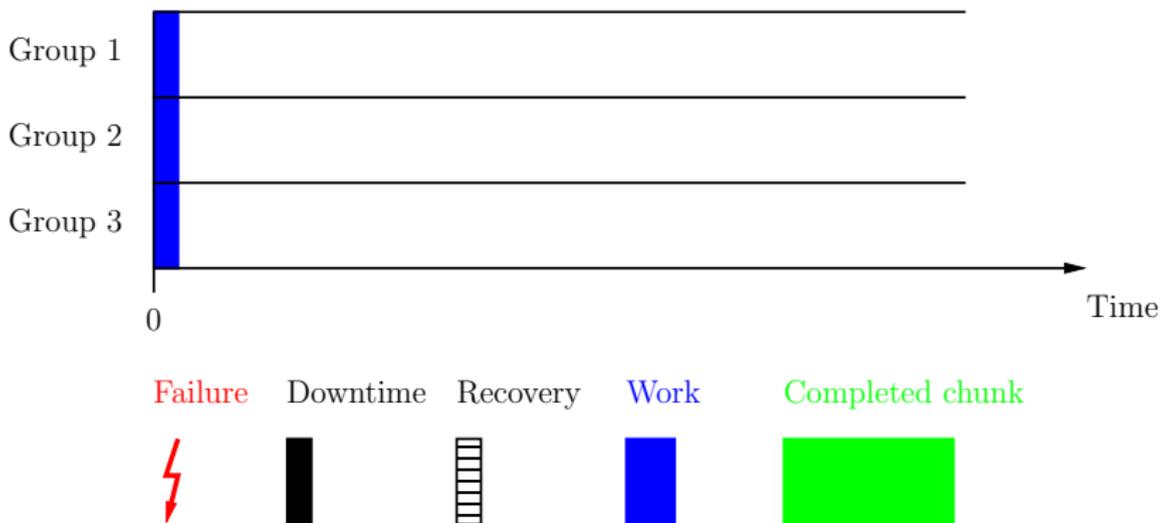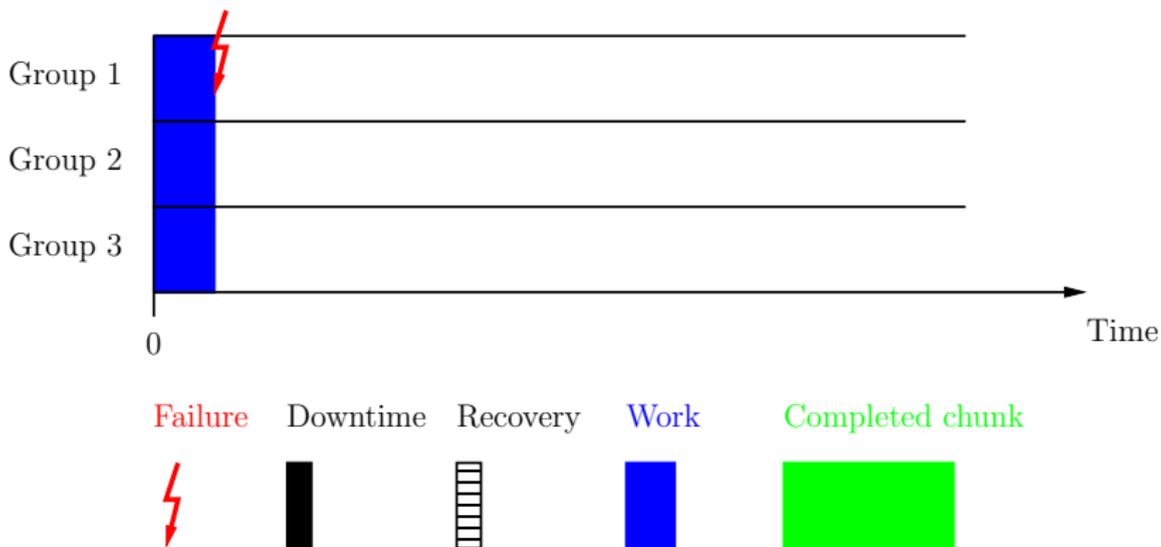
Group execution of a chunk

Group execution of a chunk

Group execution of a chunk
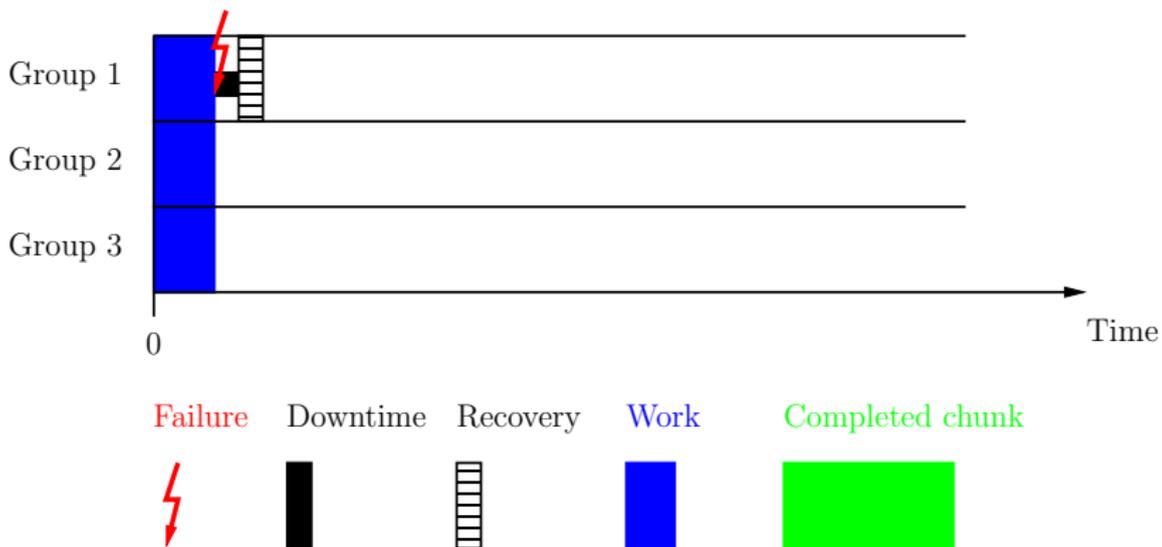
Group execution of a chunk
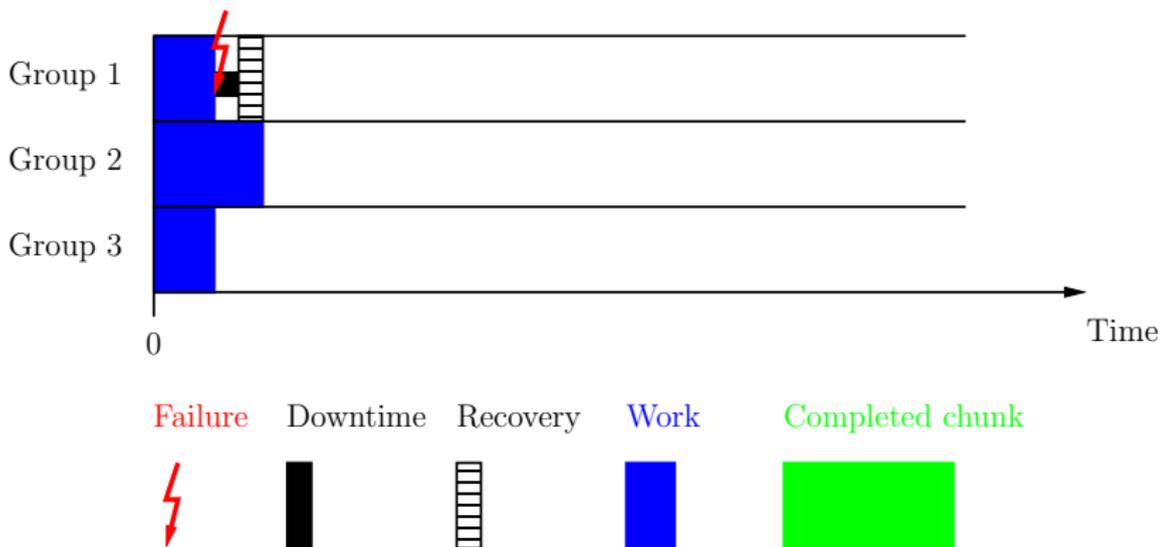
Group execution of a chunk
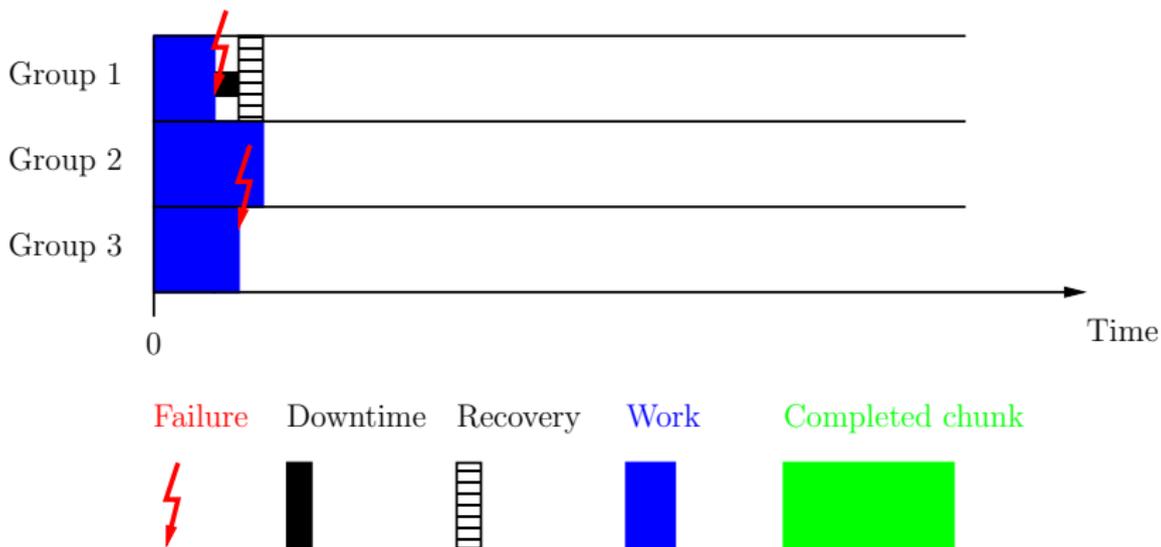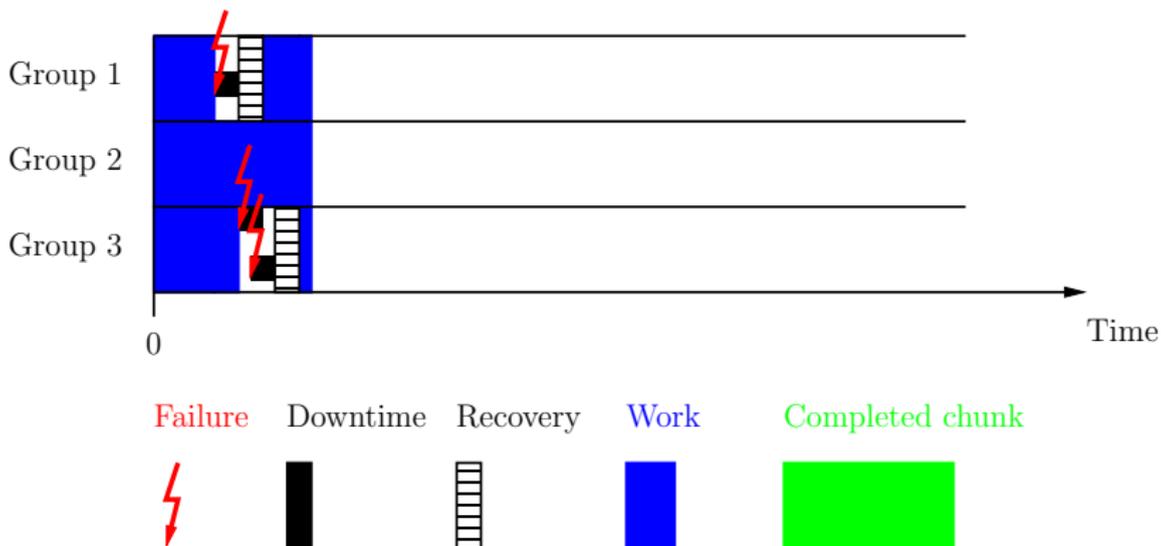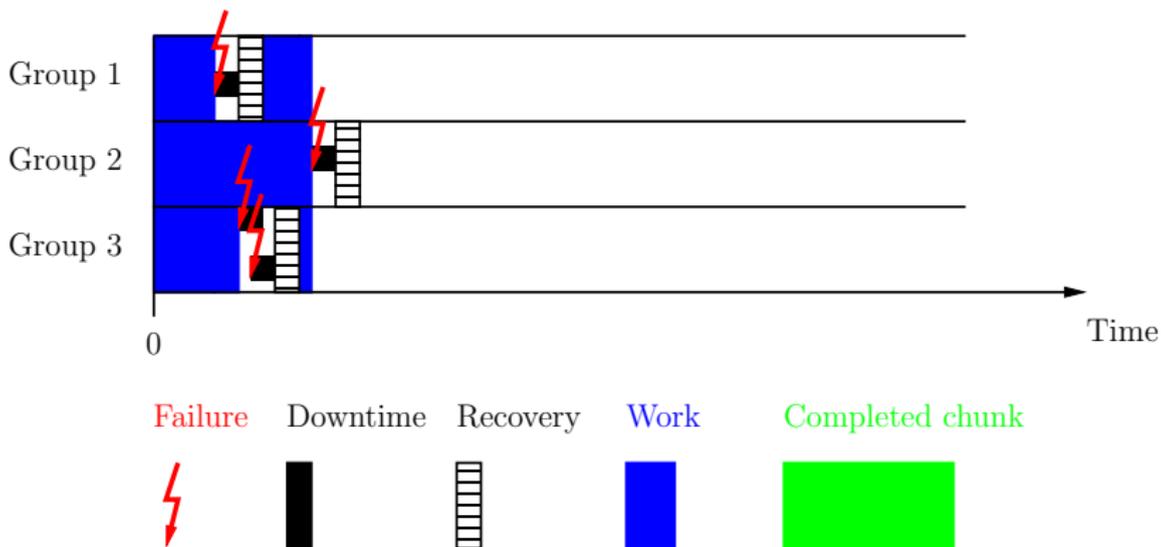
Group execution of a chunk
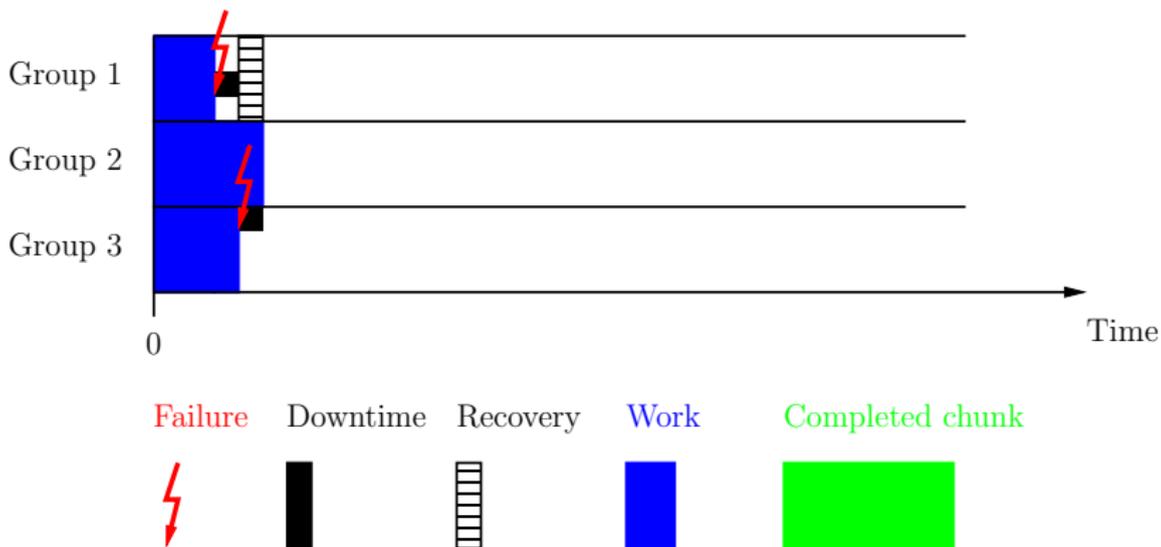
Group execution of a chunk
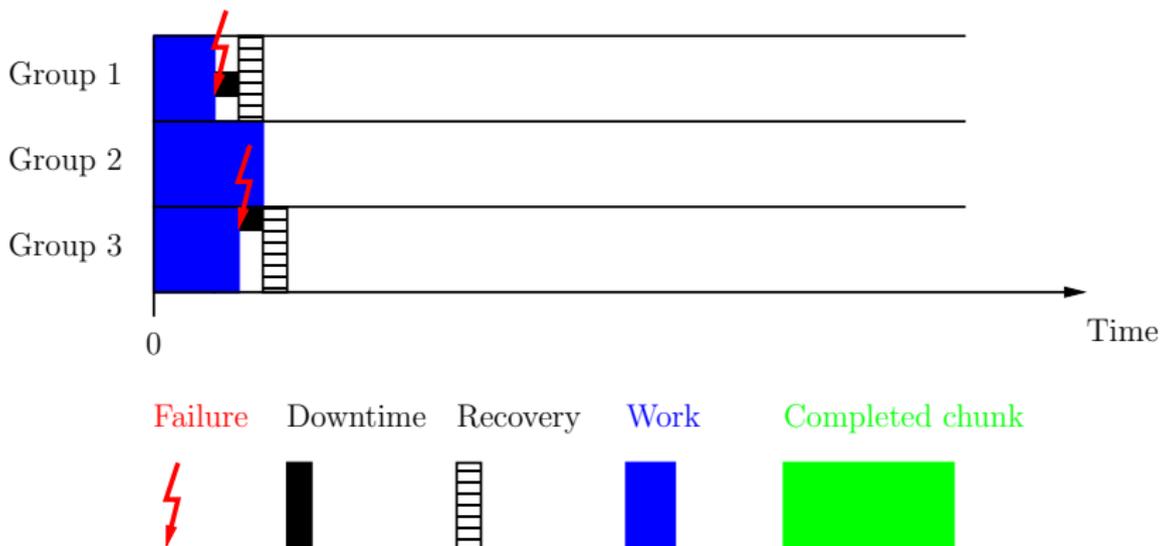
Group execution of a chunk

Group execution of a chunk

Group execution of a chunk

Group execution of a chunk

Group execution of a chunk
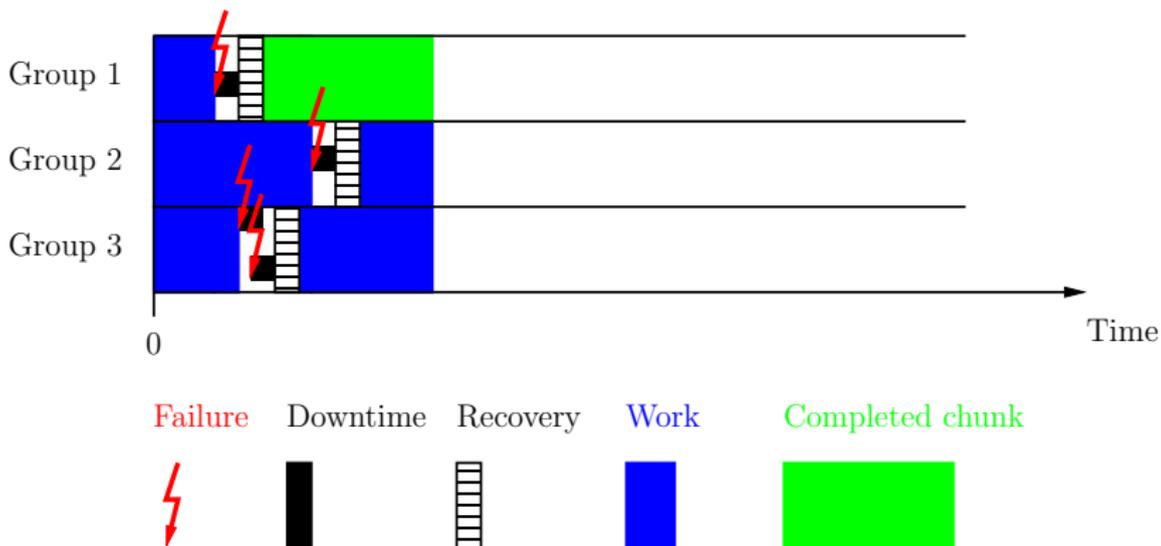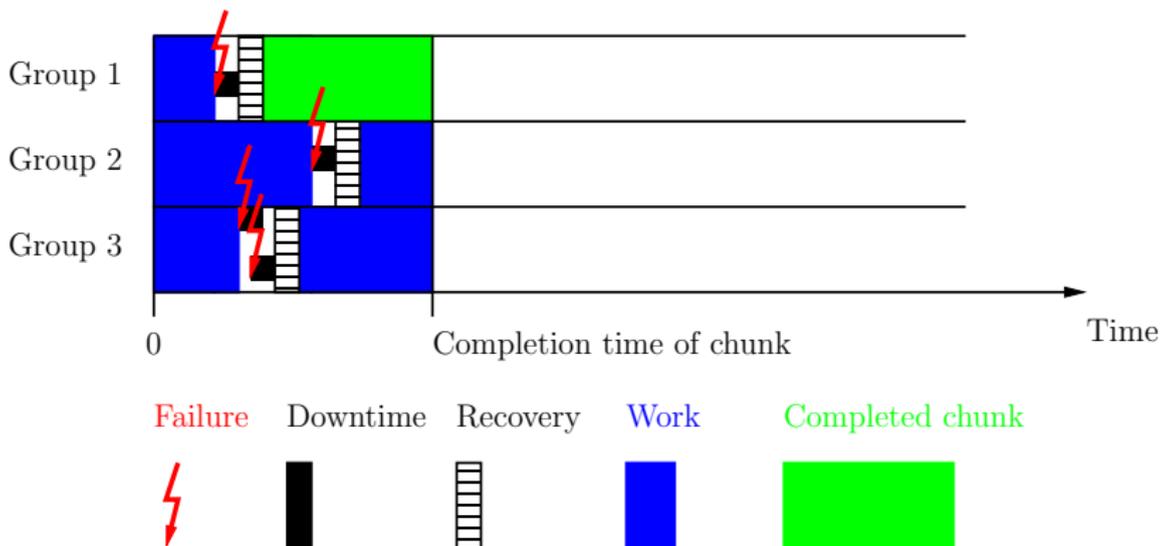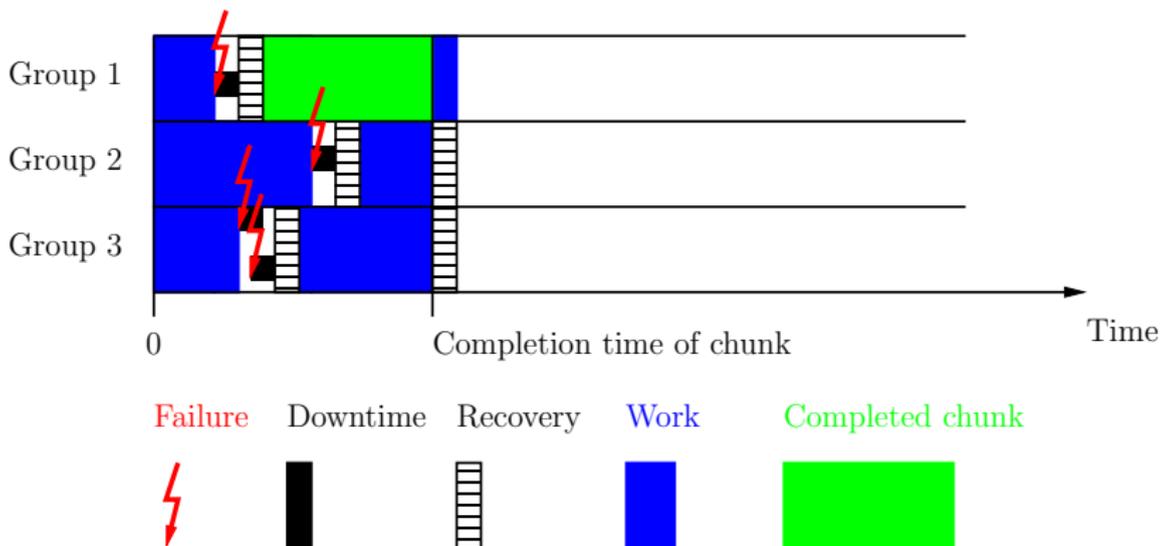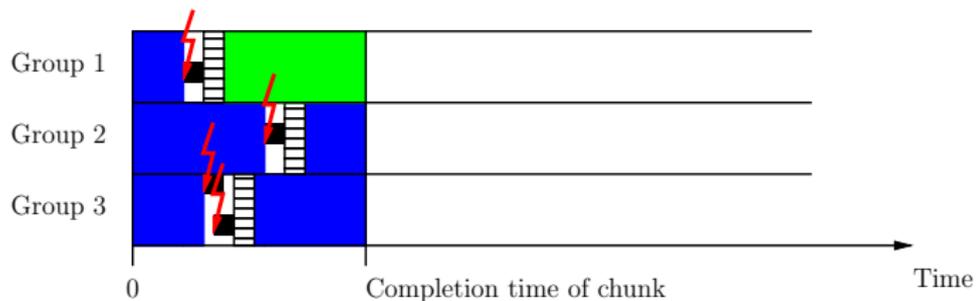
Group execution of a chunk

Group execution of a chunk

Group execution of a chunk

A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

A zest of theory

Group 1

Group 2

Group 3

0   Completion time of chunk   Time

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
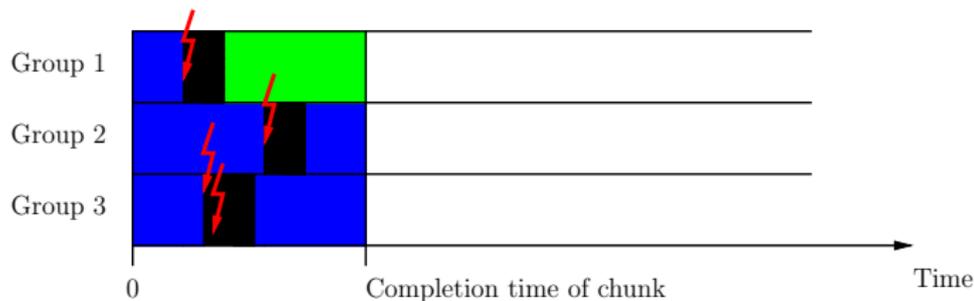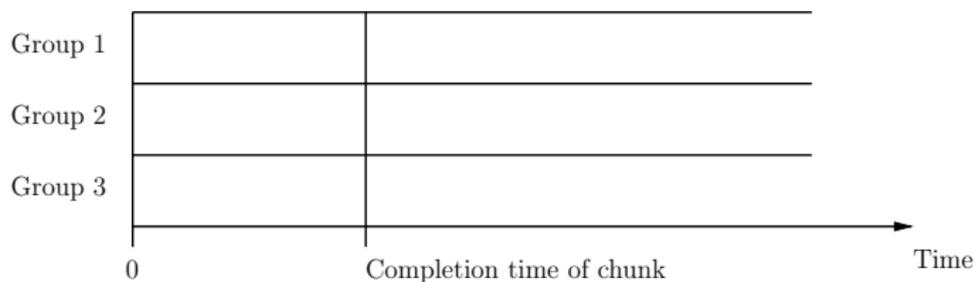
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
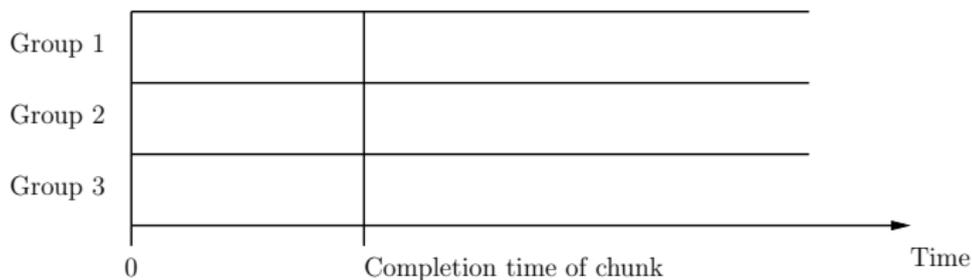
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
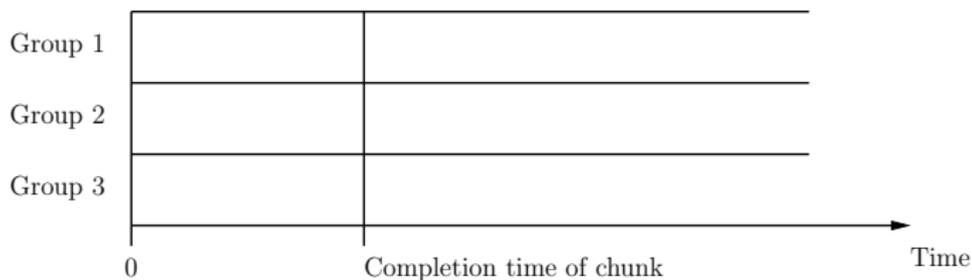
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
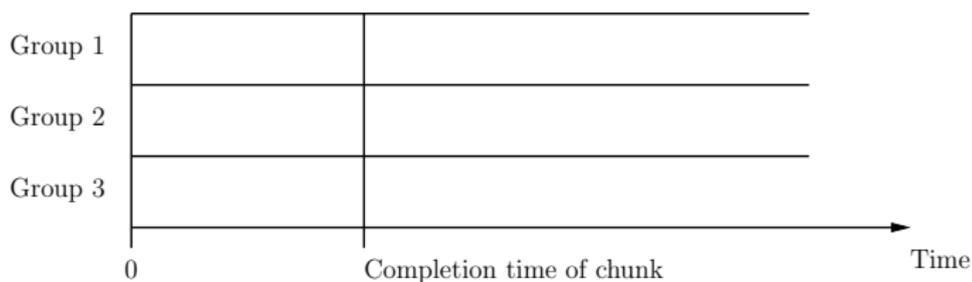
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
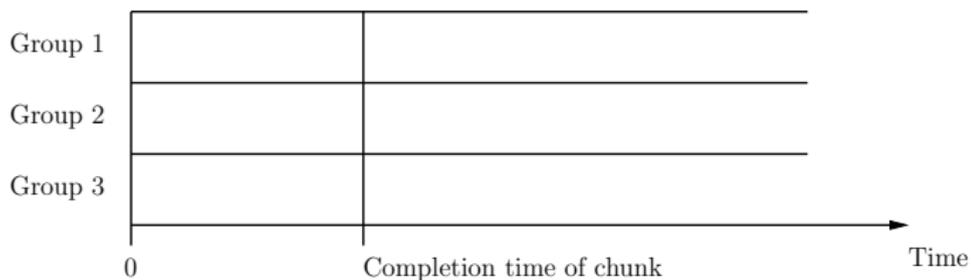
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
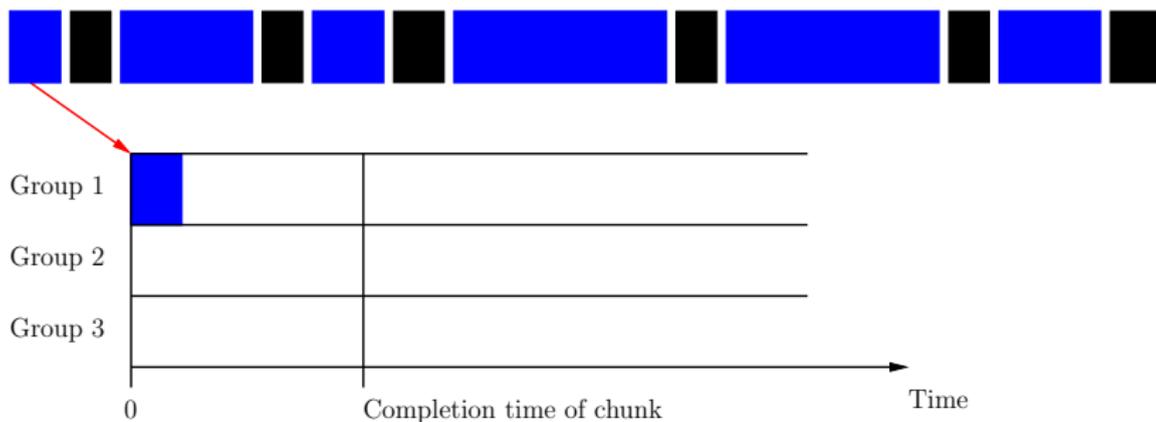
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \le \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
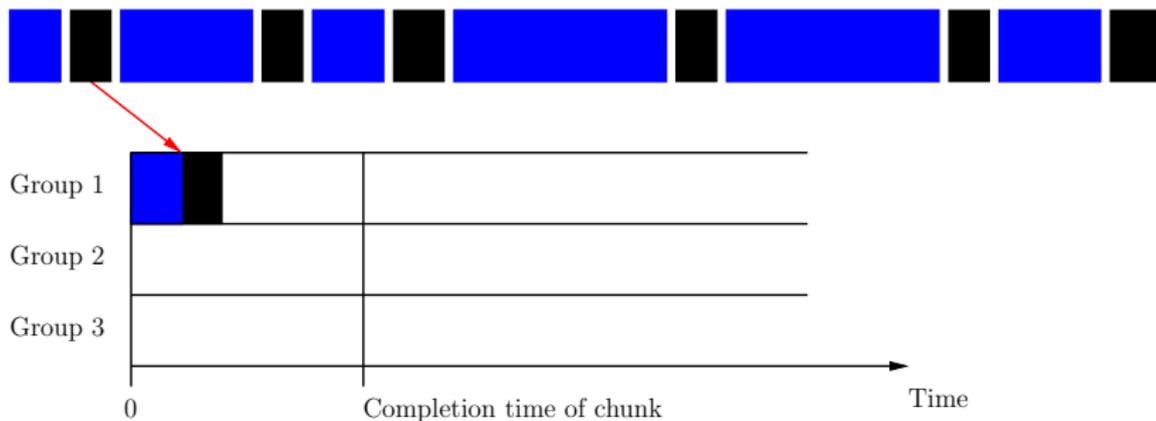
A zest of theory

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n)-1)\mathbb{E}(Y)}{g}$$

A zest of theory

$$\mathbb{E}\left(\mathcal{T}_{\mathsf{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
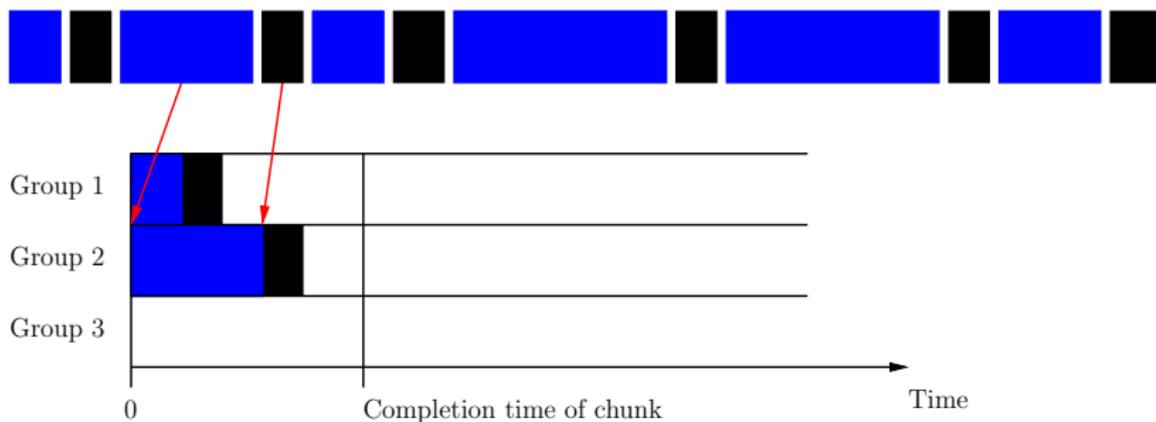
A zest of theory

Group 1

Group 2
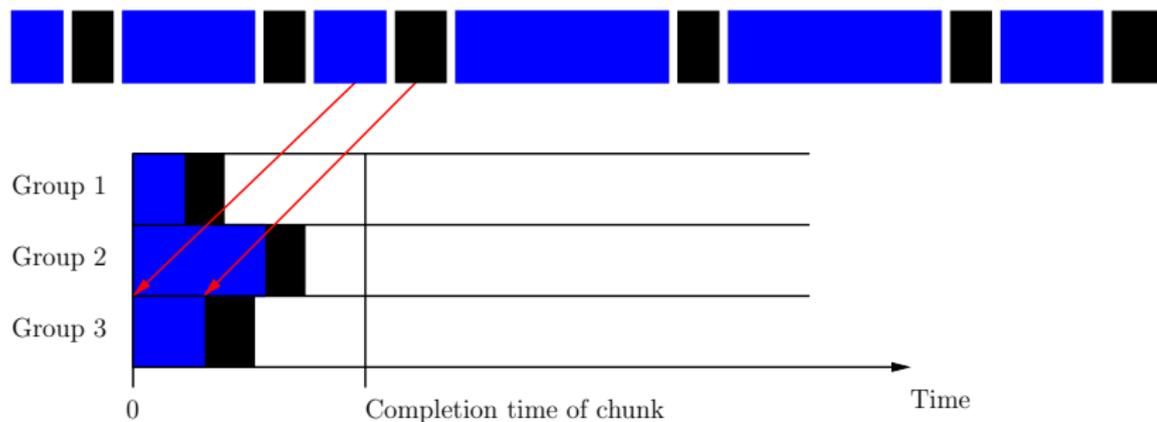
Group 3

0   Completion time of chunk   Time

$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

A zest of theory

$$\mathbb{E}\left(\mathcal{T}_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
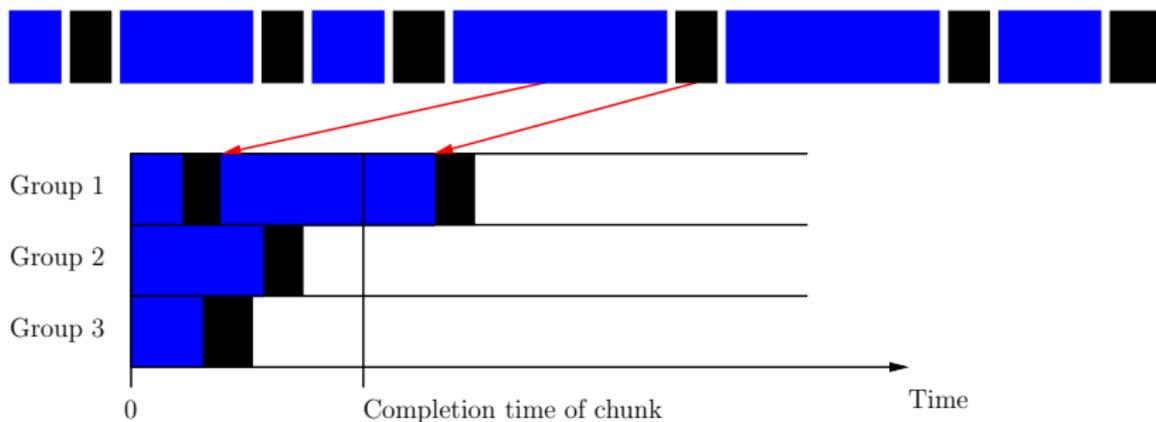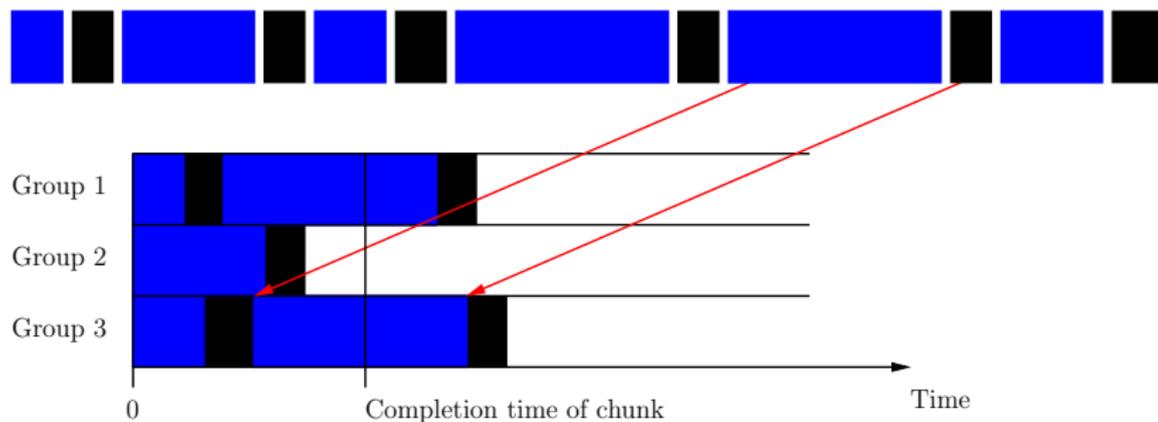
A zest of theory

Group 1

Group 2

Group 3

$0 \quad T_{\text{truestart}}^{(R+\omega_i+C)}$ Completion time of chunk   Time
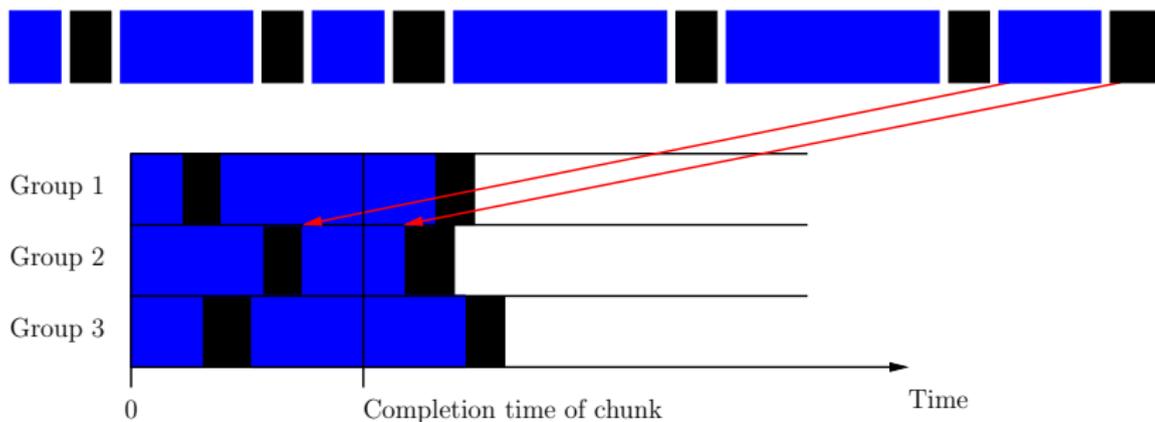
$$\mathbb{E}\left(T_{\text{truestart}}^{(R+\omega_i+C)}\right) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$
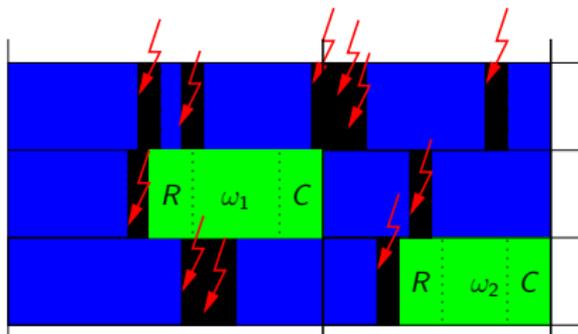
Upper bound on the expectation of the makespan (1/2)

- X: random variable for a failed attempt (blue rectangles)
- Y: random variable for the time between a failure and a full recovery (black rectangles)
- X_n is the last variable X; n is the number of attempts, including successful one (green rectangle + following blue)

Upper bound on the expectation of the makespan (1/2)

$$T_{\text{truestart}}^{(R+\omega_1+C)}$$

- X: random variable for a failed attempt (blue rectangles)
- Y: random variable for the time between a failure and a full recovery (black rectangles)
- X_n is the last variable X; n is the number of attempts, including successful one (green rectangle $+$ following blue)

Upper bound on the expectation of the makespan (2/2)

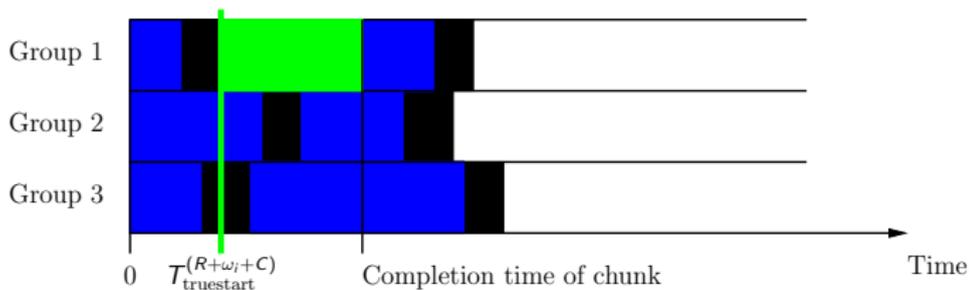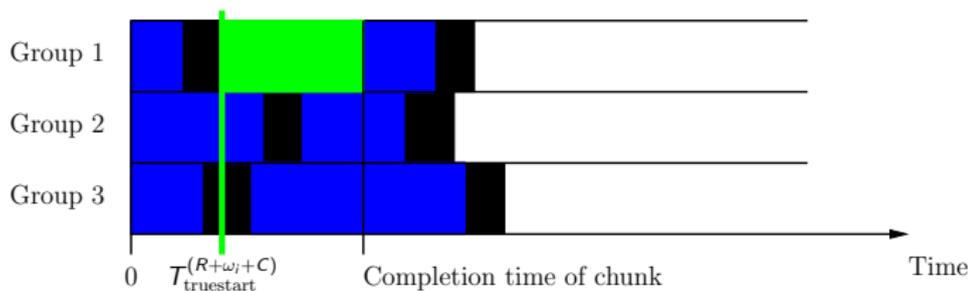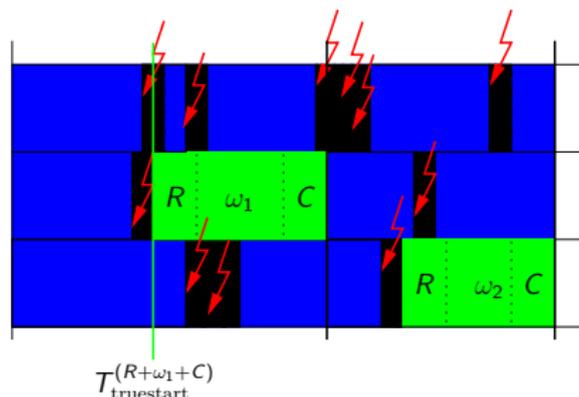$$\mathbb{E}(T_{\text{truestart}}^{(R+\omega_i+C)}) \leq \mathbb{E}(Y) + \frac{(\mathbb{E}(n)\mathbb{E}(X) - \mathbb{E}(X_n)) + (\mathbb{E}(n) - 1)\mathbb{E}(Y)}{g}$$

With $\mathbb{E}(n) = e^{\lambda p (R + \omega_i + C)}$ and $\mathbb{E}(X_n) = \frac{1}{p\lambda} + R + \omega_i + C$

Dynamic programming heuristic

- All groups use same chunk size
- Sequence of chunk sizes computed from SC'2011 paper
- LOTS of improvements in on-going work
  (different chunk size across groups, yield, ...)

Exponential distribution (MTTF $\mu = 10$ years)
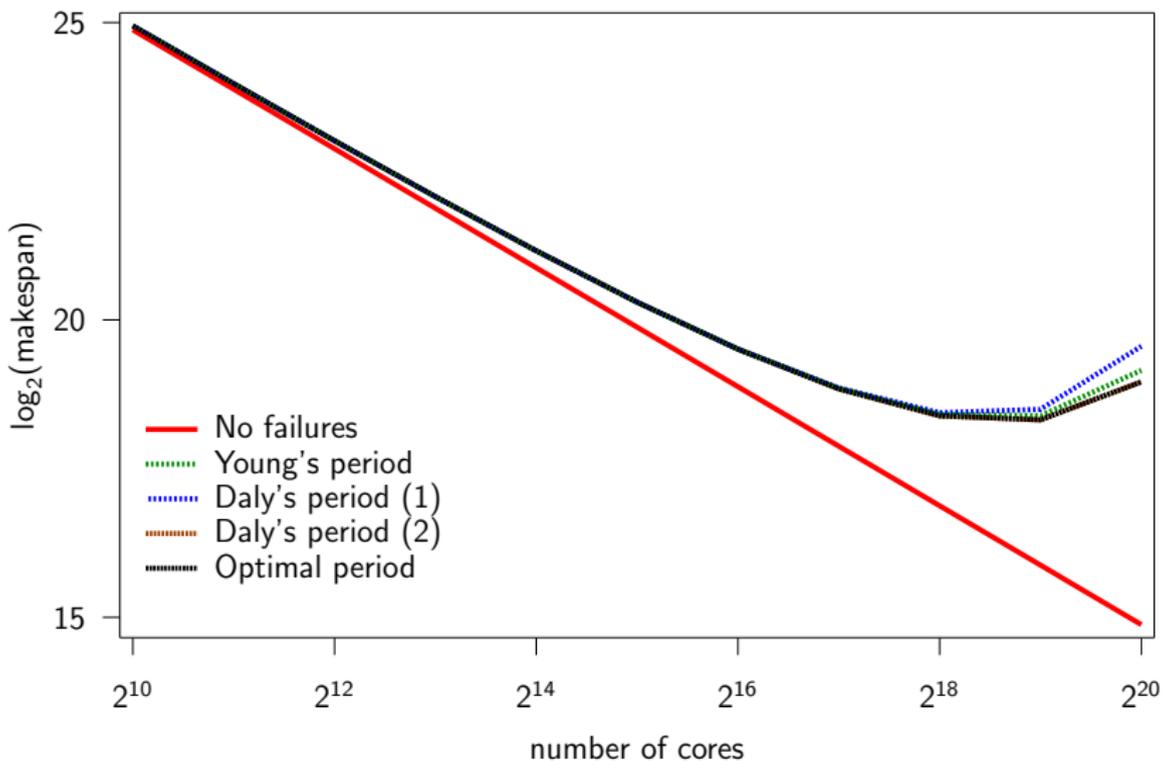
Exponential distribution (MTTF $\mu = 10$ years)
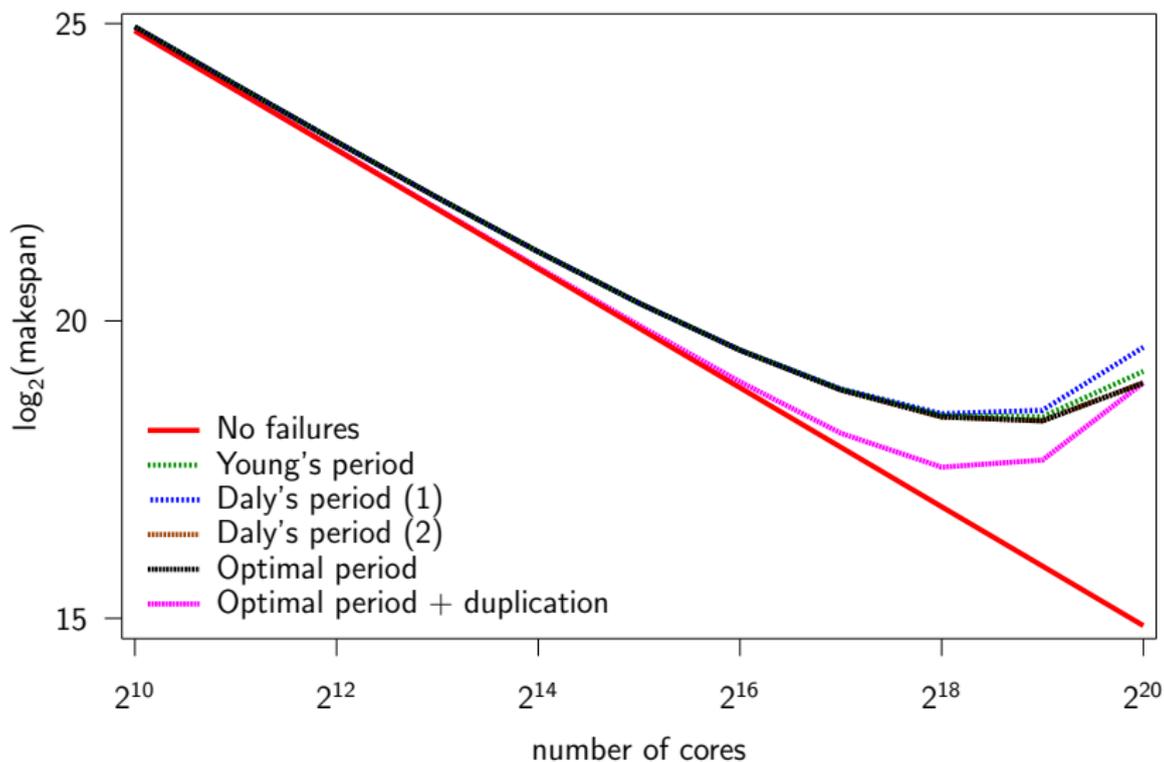
Best processor usage

Best makespan (without duplication) with 2^{19} nodes

	Without Duplication	With Duplication
Number of nodes	2^{19}	2×2^{18}
Average makespan	$344, 493$	$206, 718$

Perfect speedup with 2^{20} nodes: $\approx 32,000$

Weibull distribution (k=0.5, MTTF $\mu = 10$ years)

Legend:
- No failures
- Young's period
- Daly's period (1)
- Daly's period (2)
- Exp. optimal period
- Exp. optimal period/4
- Exp. optimal period/4 + duplication

Axes: $\log_2(\text{makespan})$ vs number of cores (2^{10} to 2^{20})
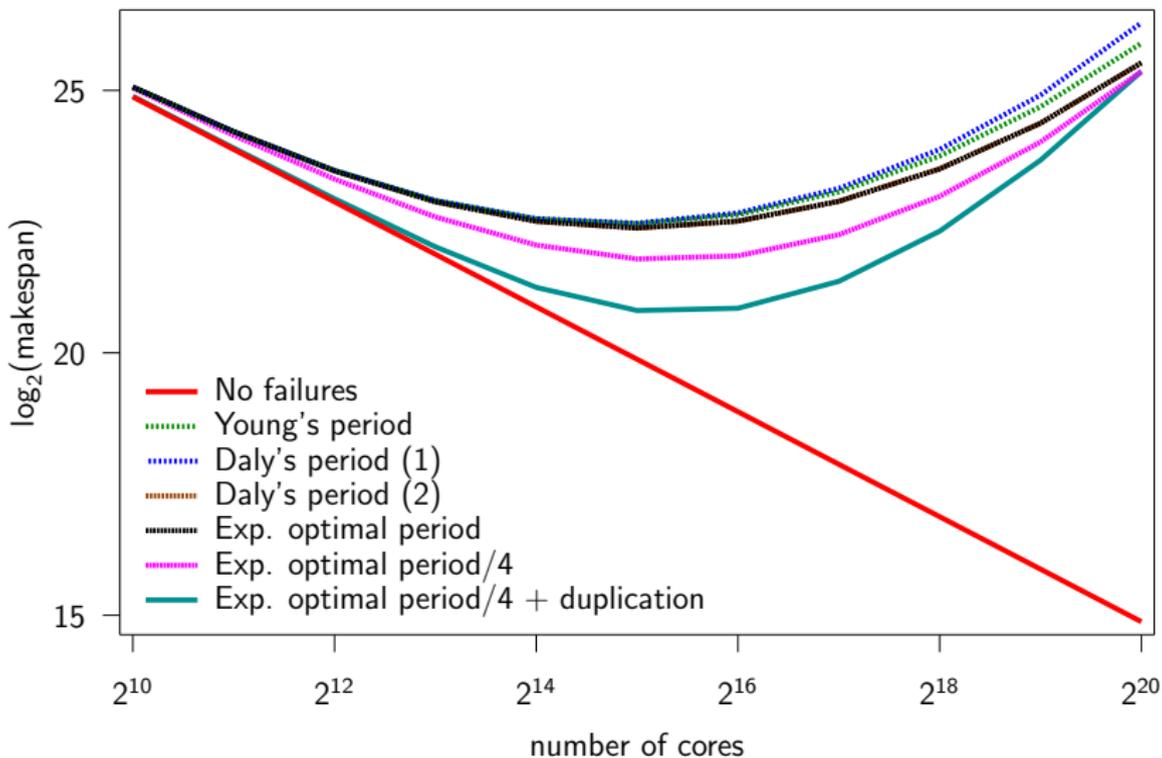
Outline

1. Best resource usage

2. Group replication

3. Process replication

4. Conclusion

PROCESS REPLICATION

- Each process replicated $g \geq 2$ times $\rightarrow$ <u>replica-group</u>
- n_{rg} = number of replica-groups ($g \times n_{rg} \leq N$)
- Study for $g = 2$ by Ferreira et al., SC'2011

Number of failures to bring down application

- $MNFTI^{\mathrm{ah}}$ Count each failure hitting any of the $g \cdot n_{rg}$ initial processors, including those already hit by a failure
- $MNFTI^{\mathrm{rp}}$ Count failures that hit running processors, and thus effectively kill replicas.

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

Number of failures to bring down application

- $MNFTI^{\mathrm{ah}}$ Count each failure hitting any of the $g \cdot n_{rg}$ initial processors, including those already hit by a failure
- $MNFTI^{\mathrm{rp}}$ Count failures that hit running processors, and thus effectively kill replicas.
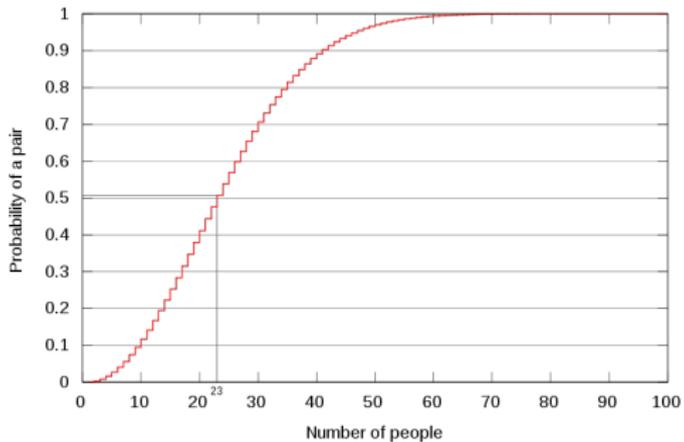
$$MNFTI^{\mathrm{ah}} = 1 + MNFTI^{\mathrm{rp}}$$

Analogy with birthday problem

Analogy with birthday problem
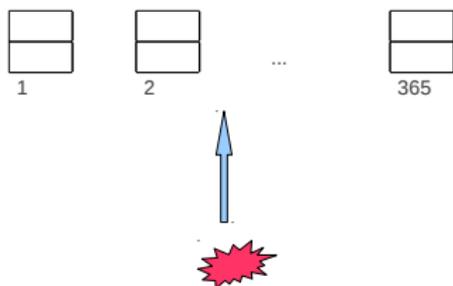
$n = n_{rg}$ bins, throw balls until one bin gets two balls

Analogy with birthday problem

$n = n_{rg}$ bins, throw balls until one bin gets two balls

Expected number of balls to throw:
$$Birthday(n) = 1 + \int_0^{+\infty} e^{-x}(1 + x/n)^{n-1}dx$$

Analogy with birthday problem
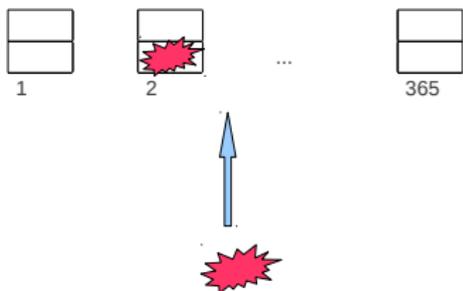
But second failure may hit already struck replica 🙁

$MNFTI^{\mathrm{ah}}$ for $g = 2$ (1/2

Theorem $MNFTI^{\mathrm{ah}} = \mathbb{E}(NFTI^{\mathrm{ah}}|0)$ where
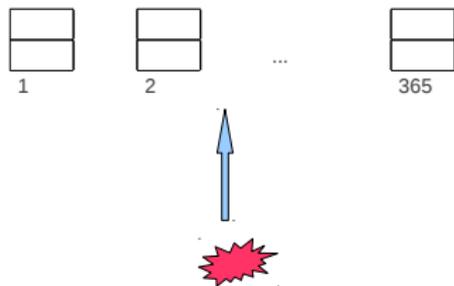
$$\mathbb{E}(NFTI^{\mathrm{ah}}|n_f) = \begin{cases} 2 & \text{if } n_f = n_{rg}, \\ \frac{2n_{rg}}{2n_{rg}-n_f} + \frac{2n_{rg}-2n_f}{2n_{rg}-n_f}\mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f+1\right) & \text{otherwise}. \end{cases}$$

$\mathbb{E}(NFTI^{\mathrm{ah}}|n_f)$: expectation of number of failures to kill application, knowing that
  • application is still running
  • failures have already hit n_f different replica-groups

$MNFTI^{\mathrm{ah}}$ for $g = 2$ (2/2)

Proof

$$\mathbb{E}\left(NFTI^{\mathrm{ah}} \,|n_{rg}\right) = \frac{1}{2} \times 1 + \frac{1}{2} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}} \,|n_{rg}\right)\right).$$

$$\mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right) = \frac{2n_{rg} - 2n_f}{2n_{rg}} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f + 1\right)\right)$$
$$+ \frac{2n_f}{2n_{rg}} \times \left(\frac{1}{2} \times 1 + \frac{1}{2}\left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right)\right)\right).$$

$MNFTI^{\mathrm{ah}}$ for $g = 2$ (2/2)

Proof

$$\mathbb{E}\left(NFTI^{\mathrm{ah}} \, |n_{rg}\right) = \frac{1}{2} \times 1 + \frac{1}{2} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}} \, |n_{rg}\right)\right).$$

$$\mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right) = \frac{2n_{rg} - 2n_f}{2n_{rg}} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f + 1\right)\right)$$
$$+ \frac{2n_f}{2n_{rg}} \times \left(\frac{1}{2} \times 1 + \frac{1}{2}\left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right)\right)\right).$$

$MNFTI^{\mathrm{ah}}$ for $g = 2$ (2/2)

> For Exponential failures:
> $MTTI = systemMTBF(2n_{rg}) \times Birthday(n_{rg})$

$$\mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right) = \frac{2n_{rg} - 2n_f}{2n_{rg}} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f + 1\right)\right)$$
$$+ \frac{2n_f}{2n_{rg}} \times \left(\frac{1}{2} \times 1 + \frac{1}{2}\left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right)\right)\right).$$

$MNFTI^{\mathrm{ah}}$ for $g = 2$ (2/2)

For Exponential failures:
$$MTTI = systemMTBF(2n_{rg}) \times \cancel{Birthday}(n_{rg})$$

$$MTTI = systemMTBF(2n_{rg}) \times MNFTI^{\mathrm{ah}}$$

$$\mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right) = \frac{2n_{rg} - 2n_f}{2n_{rg}} \times \left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f + 1\right)\right)$$
$$+ \frac{2n_f}{2n_{rg}} \times \left(\frac{1}{2} \times 1 + \frac{1}{2}\left(1 + \mathbb{E}\left(NFTI^{\mathrm{ah}}|n_f\right)\right)\right).$$

Failure distribution

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

- All replica-groups have at least one replica running
- Exponential: $R(t) = \left(1 - \left(1 - e^{-\lambda t}\right)^g\right)^{n_{rg}}$
- Weibull: $R(t) = \left(1 - \left(1 - e^{-\left(\frac{t}{\lambda}\right)^k}\right)^g\right)^{n_{rg}}$
- Can use dynamic programming algorithms from SC'2011

MTTI

- $MTTI = \int_0^{+\infty} R(t)\mathrm{d}t \rightarrow$ closed-form formulas
- Ferreira et al. significantly underestimate $MTTI$ (☹ or ☺?)

Weibull distribution (MTTF $\mu = 125$ years)

$C = R = 10mn$, $D = 1mn$, $\mathcal{W} = 10,000$ years = 3.7 parallel days)
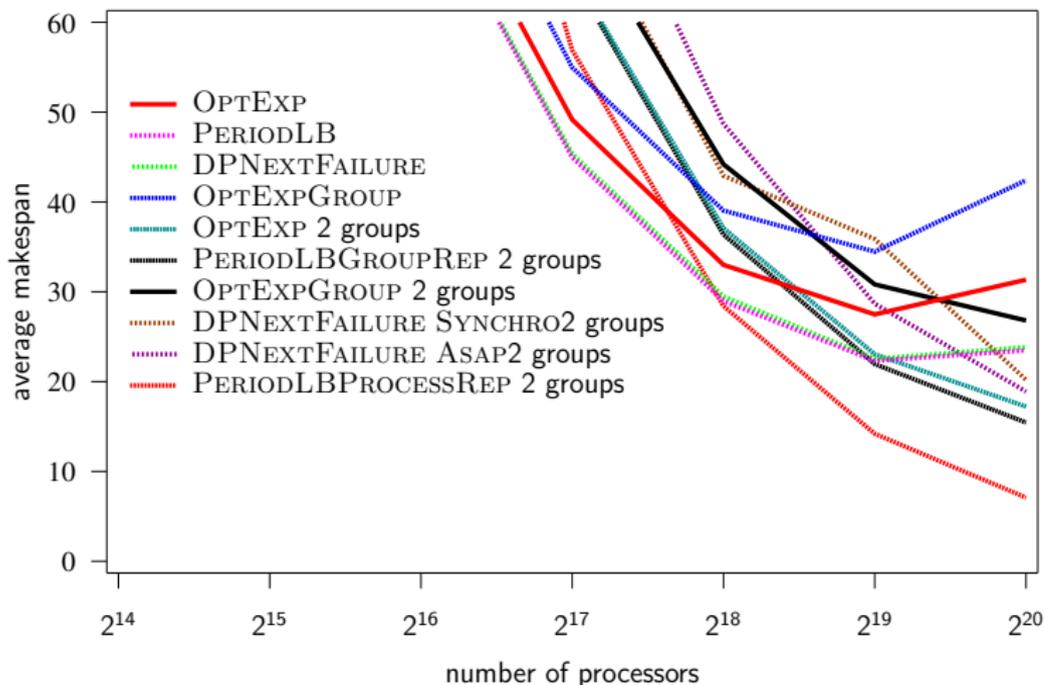
Outline

1. Best resource usage

2. Group replication

3. Process replication

4. Conclusion

Conclusion

- Software/hardware techniques to reduce checkpoint, recovery, migration times and to improve failure prediction
- Multi-criteria scheduling problem
  makespan/energy/reliability
- Best resource usage (performance trade-offs)
- Replication can improve makespan ☺

- Need combine all these approaches!

Several challenging algorithmic/scheduling problems ☺