Analysis of Large Unreliable Stochastic Networks

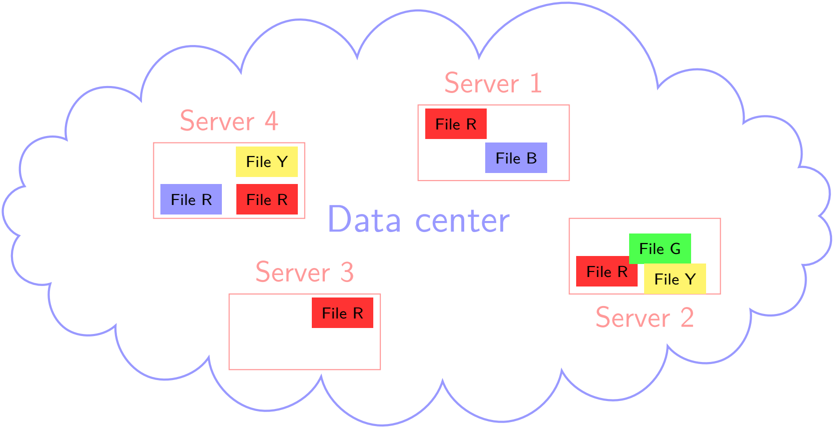
SUN Wen

(joint work with Mathieu Feuillet and Philippe Robert)

INVENTEURS DU MONDE NUMÉRIQUE

Novemeber 13, 2015, Eindhoven

Storage in data center – a simple example



- 4 servers.
- 4 files:

File R	4 copies;	File Y	2 copies;
File B	2 copies;	File G	1 copies.

- If server 2 crashed, only File G will be lost forever.

Real world: servers fail

Large data center — files are stored in servers;

10 out of 200000 — servers fail per day approximately;

Literature — Failure Trends in a Large Disk Drive Population
Eduardo Pinheiro et al.(Google, 2007)

Problem:
- How to prevent losing files?

Real world: servers fail

Large data center files are stored in servers;

10 out of 200000 servers fail per day approximately;

Literature Failure Trends in a Large Disk Drive Population
Eduardo Pinheiro et al.(Google, 2007)

Problem:

- How to prevent losing files?
  Making more copies!

Real world: servers fail

Large data center files are stored in servers;

10 out of 200000 servers fail per day approximately;

Literature Failure Trends in a Large Disk Drive Population
Eduardo Pinheiro et al.(Google, 2007)

Problem:

- How to prevent losing files?
  Making more copies!
- Effective Bandwidth Utilization?

Real world: servers fail

Large data center files are stored in servers;

10 out of 200000 servers fail per day approximately;

Literature Failure Trends in a Large Disk Drive Population
Eduardo Pinheiro et al.(Google, 2007)

Problem:

- How to prevent losing files?
  Making more copies!
- Effective Bandwidth Utilization?
  Upper limit of copies.

Literature on duplication algorithm

Experiments by data center designers OpenDHT, PAST, ...

Math models

- For single server
  An Analytical Estimation of Durability in DHTs
  F. Picconi, B. Baynat, P. Sens (2007)
- For large number of servers
  Little work has been done.

Math Model: duplication in a network with failures

Model of network with failures

Number of servers	N
Number of files	F_N
Failure rate for each server	μ
Duplication rate for each server	λ per server (global rate λN)
Maximum number of copies	d
Duplication algorithm	Depends on designers

A simplified model

Number of servers	N
Number of files	F_N
Failure rate for each copy	μ
Duplication rate for whole system	λN
Maximum number of copies	d
Duplication algorithm	Duplicate the file with least number of copies

A simplified model

Assumptions:

- Initial state: all files have d copies (maximum case).

- Scaling: average number of files per server

$$\lim_{N\to\infty} \frac{F_N}{N} = \beta.$$

Intuitive picture:

- larger d $\implies$ files are less likely lost;

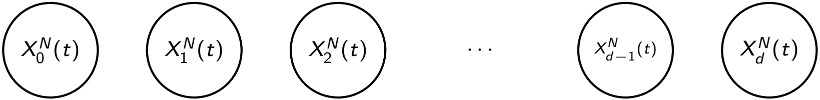
- larger β $\implies$ complete for bandwidth.

$d = 2$ A Scaling Analysis of a Transient Stochastic Network. M. Feuillet and P. Robert (2014)

Transient Markov Process: evolution among coordinates

$$X^N(t) = \left(X_0^N(t), X_1^N(t), \ldots, X_d^N(t)\right),$$

$X_i^N(t)$ = number of files with i copies at time t.



- at time 0,

$$X_d^N(0) = F_N,$$

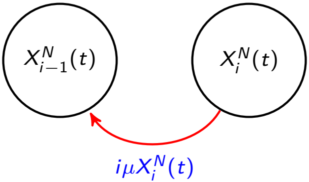
$$X_i^N(0) = 0, \qquad \forall 0 \leq i < d.$$

- $X^N(t)$ is a jump process in $\mathbb{N}^{d+1}$.

Transient Markov Process: losing a copy

$X_i^N(t)$ = number of files with i copies at time t.

evolution between coordinates



at rate $i\mu x_i$,

a file with i copies loses 1 copy, then it has $(i-1)$ copies.

$$x_i \to x_i - 1, \qquad x_{i-1} \to x_{i-1} + 1.$$

Transient Markov Process: duplication

$$X_i^N(t) = \text{number of files with } i \text{ copies at time } t.$$

evolution between coordinates

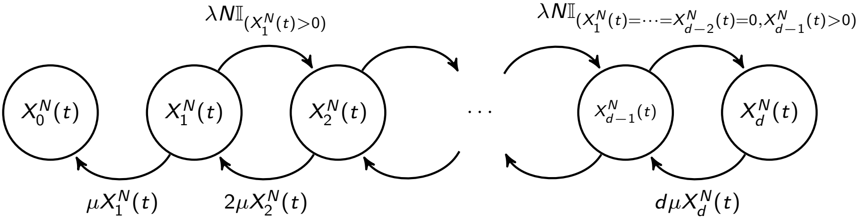
$$X_i^N(t) \xrightarrow{\lambda N \mathbb{I}_{(X_1^N(t)=\cdots=X_{i-1}^N(t)=0;\, X_i^N(t)>0)}} X_{i+1}^N(t)$$

if there is no file with less than i copies
$\left(X_1^N(t) = \cdots = X_{i-1}^N(t) = 0; X_i^N(t) > 0\right)$,
at rate λN,
a file with i copies duplicates, then it has $(i+1)$ copies.

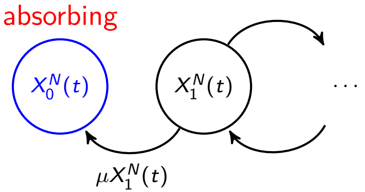
$$x_i \to x_i - 1, \qquad x_{i+1} \to x_{i+1} + 1.$$

Transient Markov Process: evolution among coordinates

$X_i^N(t)$ = number of files with i copies at time t.



Transient Markov Process: absorbing state



Absorbing state: caused by random events, with probability 1,

$$\left(X_0^N(t), X_1^N(t), \ldots, X_d^N(t)\right) \stackrel{t\to\infty}{\longrightarrow} (F_N, 0, \ldots, 0)\,.$$

All files lost eventually!

Aim: Estimate the rate of decay of the system when N is large.

Decay rate of system

For $\delta \in (0, 1)$,

$$T_N(\delta) = \text{first time for } \delta F_N \text{ files being lost}$$

$$= \inf\left\{t \geq 0 : X_0^N(t) \geq \delta F_N\right\}$$

$$\sim \inf\left\{t \geq 0 : \frac{X_0^N(t)}{N} \geq \delta\beta\right\}$$

Decay rate of system

For $\delta \in (0,1)$,

$$T_N(\delta) = \text{first time for } \delta F_N \text{ files being lost}$$

$$= \inf\left\{t \geq 0 : X_0^N(t) \geq \delta F_N\right\}$$

$$\sim \inf\left\{t \geq 0 : \frac{X_0^N(t)}{N} \geq \delta\beta\right\}$$

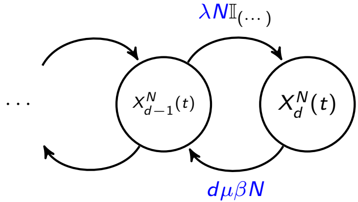
$\Longrightarrow$ Study of the scaled process $\left(\frac{X_i^N(\cdot)}{N}\right)$.

Underloaded system

$$d\mu\beta < \lambda$$

Meaning of condition: $d\mu\beta < \lambda$

At beginning , $X_d^N \sim \beta N$.



* loss rate $<$ duplication rate;
* most files stay with d copies;
* what will happen for larger time?

Underloaded system: $d\mu\beta < \lambda$

Theorem (Stability)

Initial state $\left(X_j^N(0)\right)_{j=0}^{d} = (0, \ldots, 0, F_N)$ and $\frac{F_N}{N} \to \beta$

Underloaded system: $d\mu\beta < \lambda$

Theorem (Stability)

Initial state $\quad \left(X_j^N(0)\right)_{j=0}^{d} = (0, \ldots, 0, F_N) \quad$ and $\quad \frac{F_N}{N} \to \beta$

Fluid limit $\quad \left(\frac{X_j^N(t)}{N}\right)_{j=0}^{d} \Rightarrow (0, \ldots, 0, \beta).$

Underloaded system: $d\mu\beta < \lambda$

Theorem (Stability)

Initial state $\left(X_j^N(0)\right)_{j=0}^{d} = (0, \dots, 0, F_N)$ and $\frac{F_N}{N} \to \beta$

Fluid limit $\left(\frac{X_j^N(t)}{N}\right)_{j=0}^{d} \Rightarrow (0, \dots, 0, \beta)$.

Time scale: $t \to N^{d-2}t$

$$\lim_{N\to\infty} \left(\frac{X_j^N(N^{d-2}t)}{N}\right)_{j=0}^{d} = (0, \dots, 0, \beta).$$

$\ast$ All alive files have d copies.

$\Rightarrow$ The time scale of decay is larger than $t \to N^{d-2}t$.

Underloaded system: $d\mu\beta < \lambda$

Theorem (Decay) **(Main Result!)**

Time scale: $t \to N^{d-1}t$

$$\lim_{N\to\infty} \left(\frac{X_j^N(N^{d-1}t)}{N} \right)_{j=0}^{d} = (x_0(t), 0, \ldots, 0, x_d(t))$$

Underloaded system: $d\mu\beta < \lambda$

Theorem (Decay) **(Main Result!)**

Time scale: $t \to N^{d-1}t$

$$\lim_{N\to\infty}\left(\frac{X_j^N(N^{d-1}t)}{N}\right)_{j=0}^{d} = (x_0(t), 0, \dots, 0, x_d(t))$$

$$\begin{cases} x_0(t) + x_d(t) \equiv \beta, \\ \mathrm{d}x_0(t) = \lambda\frac{(d-1)!}{\rho^{d-1}}\frac{d\mu(\beta - x_0(t))}{\lambda - d\mu(\beta - x_0(t))}\mathrm{d}t \end{cases}$$

- Most alive files have d copies.
- Eventually, all the files will be lost.

Underloaded system: $d\mu\beta < \lambda$

Sketch of proof (Decay: $t \to N^{d-1}t$)

Step. 1 Prove for all $k = 1, 2, \ldots, d-1$,

$$\lim_{N\to\infty} \left(\frac{X_k^N(N^{d-1}t)}{N} \right) = 0.$$

Step. 2 Obtain the limit of

$$\left(\frac{X_0^N(N^{d-1}t)}{N} \right) = \left(\mu \frac{1}{N} \int_0^{N^{d-1}t} X_1^N(u)\mathrm{d}u + \text{martingale part} \right).$$

Underloaded system: $d\mu\beta < \lambda$

Sketch of proof (Decay: $t \to N^{d-1}t$)

Step. 1 Prove for all $k = 1, 2, \ldots, d-1$,

$$\lim_{N\to\infty}\left(\frac{X_k^N(N^{d-1}t)}{N}\right) = 0.$$

Underloaded system: $d\mu\beta < \lambda$

Sketch of proof (Decay: $t \to N^{d-1}t$)

Step. 1 Prove for all $k = 1, 2, \dots, d-1$,

$$\lim_{N\to\infty} \left(\frac{X_k^N(N^{d-1}t)}{N} \right) = 0.$$

Proof $Z^N(t) = (d-1)X_1^N(t) + (d-2)X_2^N(t) + \cdots + X_{d-1}^N(t).$

Coupling $Z^N(t) \leq L(Nt),$

$L(t) =$ an ergodic M/M/1 queue

with $+1$ rate $d\mu\beta$, -1 rate λ.

$$\sup_{t\leq T} \frac{X_k^N(N^{d-1}t)}{N} \leq \sup_{t\leq T} \frac{Z(N^{d-1}t)}{N} \leq \sup_{t\leq T} \frac{L(N^d t)}{N} \to 0.$$

Underloaded system: $d\mu\beta < \lambda$

Sketch of proof (Decay: $t \to N^{d-1}t$)

Step. 2 Obtain the limit of

$$\left(\frac{X_0^N(N^{d-1}t)}{N}\right) = \left(\mu \int_0^{N^{d-1}t} \frac{X_1^N(u)}{N} \mathrm{d}u + \text{martingale part}\right).$$

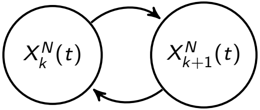
Underloaded system: $d\mu\beta < \lambda$

Sketch of proof (Decay: $t \to N^{d-1}t$)

Step. 2 Obtain the limit of

$$\left(\frac{X_0^N(N^{d-1}t)}{N}\right) = \left(\mu \int_0^{N^{d-1}t} \frac{X_1^N(u)}{N} \mathrm{d}u + \text{martingale part}\right).$$

Balance of flows **(Key point!)**



$$\lim_{N\to\infty} \frac{1}{N^{k+1}} \left(\int_0^{N^{d-1}t} \left[\mu(k+1)X_{k+1}^N(u) - \lambda N X_k^N(u)\right] \mathrm{d}u\right) = 0.$$

$$\Longrightarrow \lim_{N\to\infty} \left(\int_0^{N^{d-1}t} \frac{X_1^N(u)}{N} - \frac{(d-1)!}{\rho^{d-2}} \frac{X_{d-1}^N(u)}{N^{d-1}} \mathrm{d}u\right) = 0.$$

(Stochastic Calculus with Poisson Processes)

Underloaded system: $d\mu\beta < \lambda$

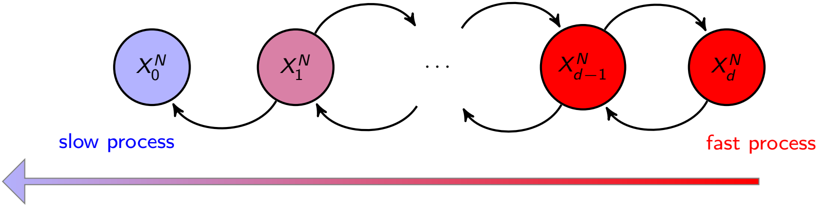
Sketch of proof (Decay: $t \to N^{d-1}t$)

Tightness and convergence of

$$\left(\int_0^{N^{d-1}t} \frac{X_{d-1}^N(u)}{N^{d-1}}\mathrm{d}u\right) = \left(\int_0^t X_{d-1}^N(N^{d-1}u)\mathrm{d}u\right)$$

Stochastic averaging problem

Markov process under time scale: $t \to N^{d-1}t$

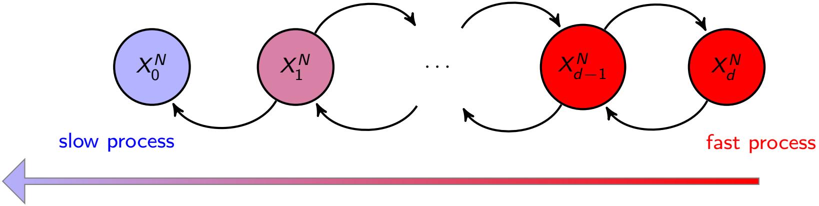


Stochastic averaging problem: Literature

- PDE: Singular Perturbation Theory;
- Probability
  - Khasminskii (1966), Freidlin, Wentzell (1979),
  - Papanicolaou, Stroock, Varadhan (1977) for diffusions,
  - Kurtz (1992) for jump processes.

Underloaded system: $d\mu\beta < \lambda$ Sketch of proof

Markov process under time scale: $t \to N^{d-1}t$

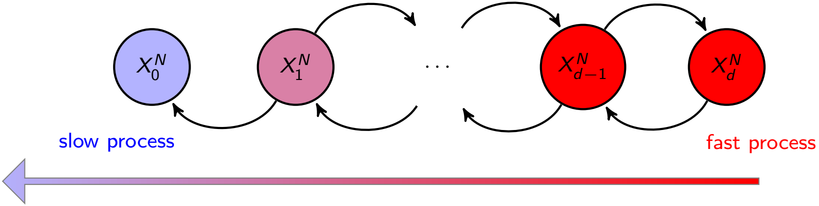


- Tightness of measures (π^N) on $\mathbb{N} \times \mathbb{R}^+$ defined by

$$< \pi^N, g >= \int_{\mathbb{R}^+} g(X_{d-1}^N(N^{d-1}u), u)\mathrm{d}u.$$

Underloaded system: $d\mu\beta < \lambda$ Sketch of proof

Markov process under time scale: $t \to N^{d-1}t$



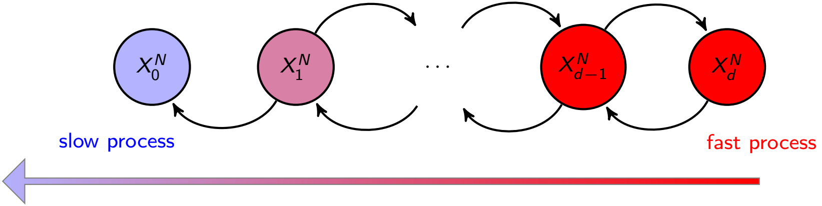
- Tightness of measures (π^N) on $\mathbb{N} \times \mathbb{R}^+$ defined by

$$< \pi^N, g >= \int_{\mathbb{R}^+} g(X_{d-1}^N(N^{d-1}u), u)\mathrm{d}u.$$

- $\pi^N \Rightarrow \pi$ with $\pi(\mathbb{N} \times [0, t]) = t$.

Underloaded system: $d\mu\beta < \lambda$ Sketch of proof

Markov process under time scale: $t \to N^{d-1}t$



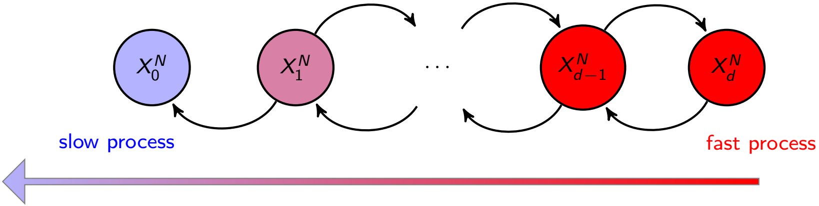
- Tightness of measures (π^N) on $\mathbb{N} \times \mathbb{R}^+$ defined by

$$< \pi^N, g >= \int_{\mathbb{R}^+} g(X_{d-1}^N(N^{d-1}u), u)\mathrm{d}u.$$

- $\pi^N \Rightarrow \pi$ with $\pi(\mathbb{N} \times [0,t]) = t$.
- Identify π: for fixed t,
  - $X_{d-1}^N(N^{d-1}t + \frac{u}{N}) \sim$ an ergodic M/M/1 queue with $+1$ rate $d\mu x_d(t)$ and -1 rate λ
  - $\pi(\cdot, t)$ is the invariant measure.

Underloaded system: $d\mu\beta < \lambda$ Sketch of proof

Markov process under time scale: $t \to N^{d-1}t$



$\pi(\cdot, t)$ is the invariant measure of an ergodic $M/M/1$ queue with $+1$ rate $d\mu x_d(t)$ and -1 rate λ.

$$\Longrightarrow x_0(t) = \lim_{N\to\infty} \lambda \frac{(d-1)!}{\rho^{d-1}} \int_0^t X_{d-1}^N(N^{d-1}u)\mathrm{d}u$$

$$= \lambda \frac{(d-1)!}{\rho^{d-1}} \int_0^t < \pi(\cdot, u), I > \mathrm{d}u$$

$$= \lambda \frac{(d-1)!}{\rho^{d-1}} \int_0^t \frac{d\mu x_d(u)}{\lambda - d\mu x_d(u)} \mathrm{d}u,$$

$$x_d(t) = \beta - x_0(t).$$

Underloaded system: $d\mu\beta < \lambda$

Corollary (Decay time for a fraction δ of files lost)

$$\forall \delta \in (0,1),\ T_N(\delta) = \inf\left\{ t \geq 0 : \frac{X_0^N(t)}{N} \geq \delta\beta \right\}$$

Underloaded system: $d\mu\beta < \lambda$

Corollary (Decay time for a fraction δ of files lost)

$$\forall \delta \in (0,1),\ T_N(\delta) = \inf\left\{ t \geq 0 : \frac{X_0^N(t)}{N} \geq \delta\beta \right\}$$

we have already proved

$$\left(\frac{X_0^N(N^{d-1}t)}{N}\right) \Longrightarrow (x_0(t)),$$

Underloaded system: $d\mu\beta < \lambda$

Corollary (Decay time for a fraction δ of files lost)

$$\forall \delta \in (0,1),\ T_N(\delta) = \inf\left\{ t \geq 0 : \frac{X_0^N(t)}{N} \geq \delta\beta \right\}$$

we have already proved

$$\left(\frac{X_0^N(N^{d-1}t)}{N} \right) \Longrightarrow (x_0(t)),$$

then for the convergence of distribution

$$\Longrightarrow \qquad \lim_{N\to\infty} \frac{T_N(\delta)}{N^{d-1}} = \frac{\rho^{d-1}}{\lambda(d-1)!}\left(-\frac{\rho}{d}\log(1-\delta) - \beta\delta\right).$$

$*$ The order of decay time is $O(N^{d-1})$.

Central limit theorem

(for stochastic averaging problem)

Central limit theorem for stochastic averaging problem

- Law of large numbers:

$$\left(\frac{X_0^N(N^{d-1}t)}{N}\right) \to (x_0(t)).$$

- **Theorem CLT** (Underloaded case)

$$\left(\frac{X_0^N(N^{d-1}t) - x_0(t)N}{\sqrt{N}}\right) \to (W(t)),$$

where $W(t)$ satisfies a *SDE*

$$\mathrm{d}W(t) = \sqrt{x_0'(t)}\mathrm{d}\mathcal{B}(t) - \frac{\lambda^2 \mu d!}{\rho^{d-1}} \frac{W(t)}{(\lambda - d\mu(\beta - x_0(t)))^2}\mathrm{d}t$$

($\mathcal{B}(t)$ is the standard Brownian motion).

- Technical point: "refined" balance of flow.

Overloaded system

$$\lambda < d\beta\mu$$

Overloaded system – $\lambda < d\beta\mu$

$p = \lfloor \frac{\lambda}{\beta\mu} \rfloor$, $\rho = \frac{\lambda}{\mu}$.

Fluid limit $\quad (\frac{X_j^N(t)}{N})_{j=0}^d \Rightarrow (x_j(t))_{j=0}^d$.

Technical point: Generalized Skorokhod Problem

$$(x_p, x_{p+1}, (x_j)_{j \neq p,p+1})(t) \xrightarrow{t\to\infty} ((p+1)\beta - \rho, \rho - p\beta, (0))$$

Overloaded system – $\lambda < d\beta\mu$

$p = \lfloor \frac{\lambda}{\beta\mu} \rfloor$, $\rho = \frac{\lambda}{\mu}$.

Fluid limit $\quad (\frac{X_j^N(t)}{N})_{j=0}^d \Rightarrow (x_j(t))_{j=0}^d$.

Technical point: Generalized Skorokhod Problem

$$(x_p, x_{p+1}, (x_j)_{j \neq p,p+1})(t) \xrightarrow{t\to\infty} ((p+1)\beta - \rho, \rho - p\beta, (0))$$

Time scale: $t \to N^{p-1}t$

$$\lim_{N\to\infty} \left(\frac{X_j^N(N^{p-1}t)}{N} \right) = (x_0(t), 0, \ldots, 0, x_p(t), x_{p+1}(t), 0, \ldots, 0)$$

Technical point: Coupling + Stochastic Averaging Problem.

Conclusion

Conclusion

Underloaded case: $\lambda > d\beta\mu$

- Fluid limit at normal time.
- Fluid limit at time scale $t \to N^{d-1}t$.
- Central limit theorem.

Overloaded case: $p\beta\mu \leq \lambda < (p+1)\beta\mu$

- Fluid limit at normal time.
- Local equibrium within p or $p+1$ copies.
- Fluid limit at time scale $t \to N^{p-1}t$.

T_δ = time for δF_N files being lost

If $(1-\delta)\beta \in \left(\frac{\lambda}{(p+1)\mu}, \frac{\lambda}{p\mu}\right)$, then $T_\delta \sim O(N^{p-1})$.

More details

Analysis of large unreliable stochastic networks

W. Sun, M. Feuillet and P. Robert (2015, arXiv)

Further work

- failure rate per copy $\longrightarrow$ failure rate per server — μ,
- copy rate of whole system — λN
  $\longrightarrow$ copy rate per server — λ.
- mean field approach.

  Ongoing work with Reza Aghajani (Brown University)
  and Philippe Robert (INRIA).

Thank you!