

P2P storage systems: towards a more accurate model

Abdulhalim Dandoush

INRIA Sophia Antipolis

MAESTRO Team

Joint work with Sara Alouf and Philippe Nain

9^{ème} atelier, June 1-4, 2008

Motivation

■ Le modèle Client / Serveur

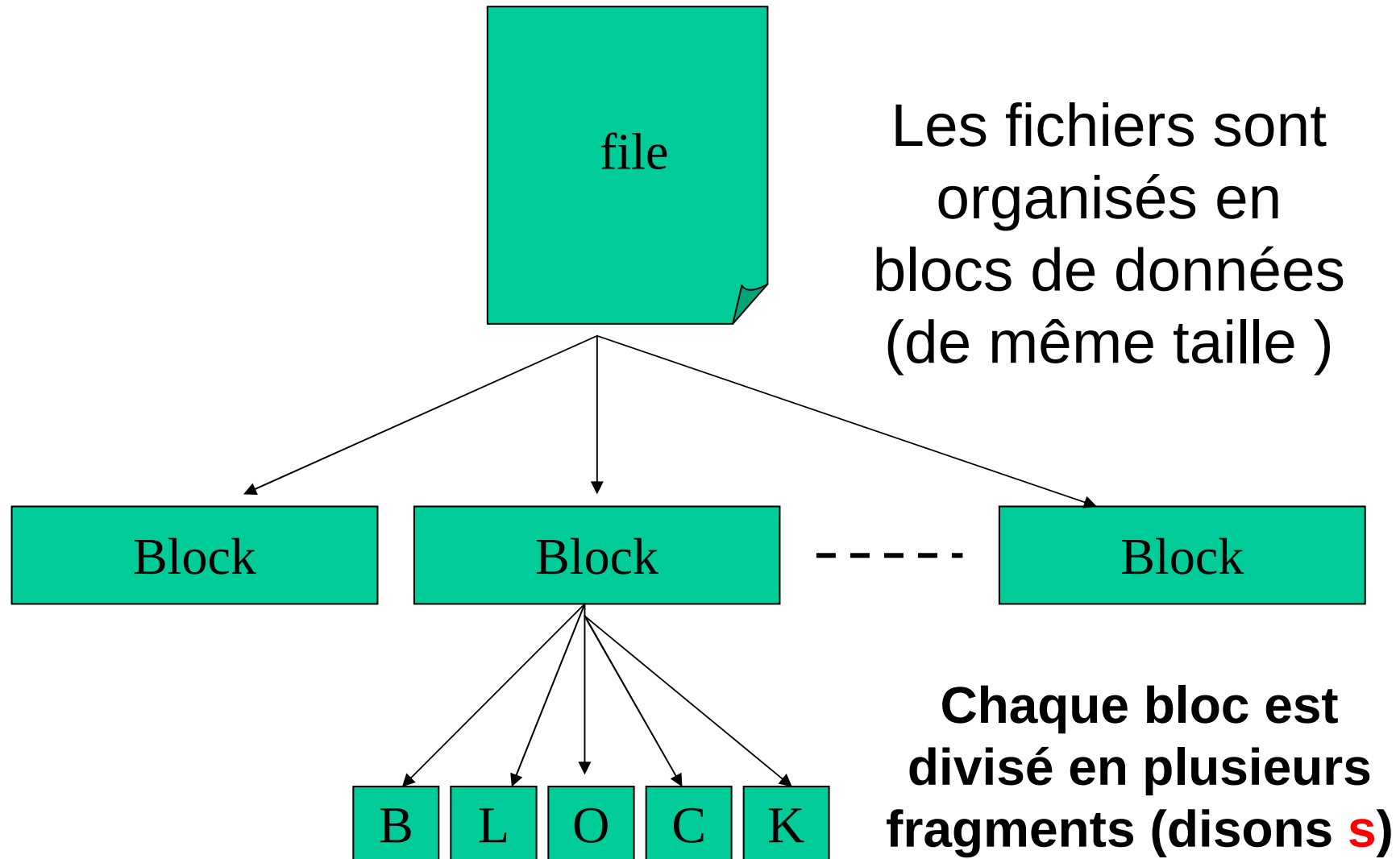
- ▶ s'appuie sur de puissants et robustes serveurs dédiés pour le stockage
- ▶ solution très fiable, simple, facile à administrer
- ▶ très coûteuse, **ne passe pas à l'échelle**

■ solution alternative: utiliser des pairs peu coûteux (infrastructures p2p), **passe très bien à l'échelle**

■ **pose** plusieurs problèmes de :

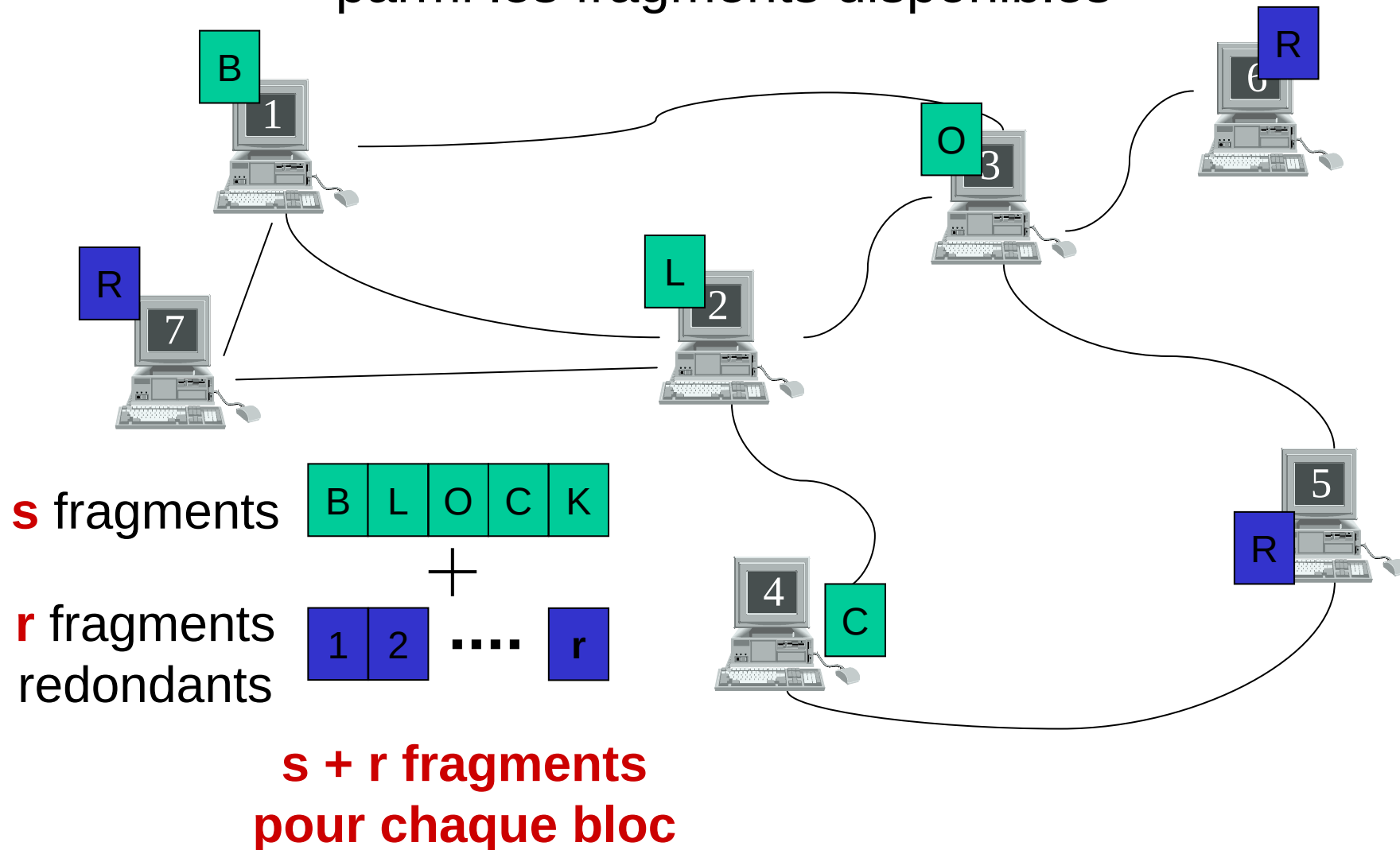
- ▶ **fiabilité**
- ▶ **disponibilité**
- ▶ localisation efficace des données
- ▶ sécurité et confidentialité

Comment peut-on stocker dans un système P2P?



La redondance augmente la fiabilité

reconstruction = n'importe quels **s**
parmi les fragments disponibles



La redondance sans réparation n'est pas efficace

déconnexion
temporaire

panne



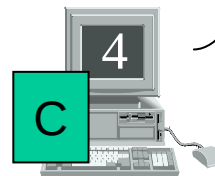
Nous devons reconstruire les fragments perdus sur
de nouveaux pairs afin de ne pas perdre
définitivement le bloc

→ 2 mécanismes de recouvrement (réparation)

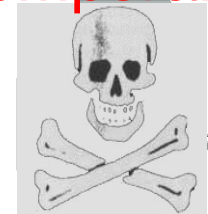


moins de s fragments

→ **bloc perdu !**

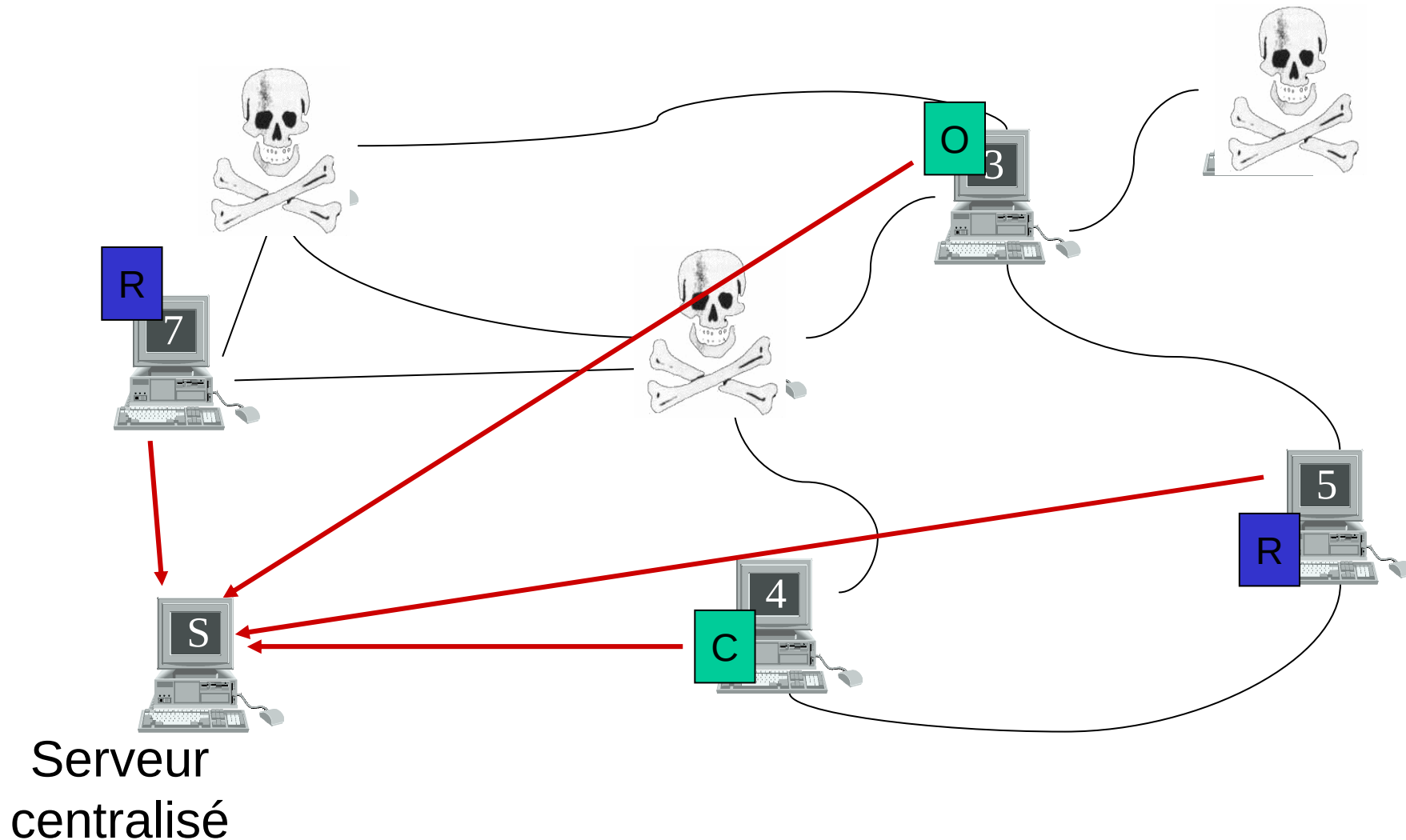


on
temporaire

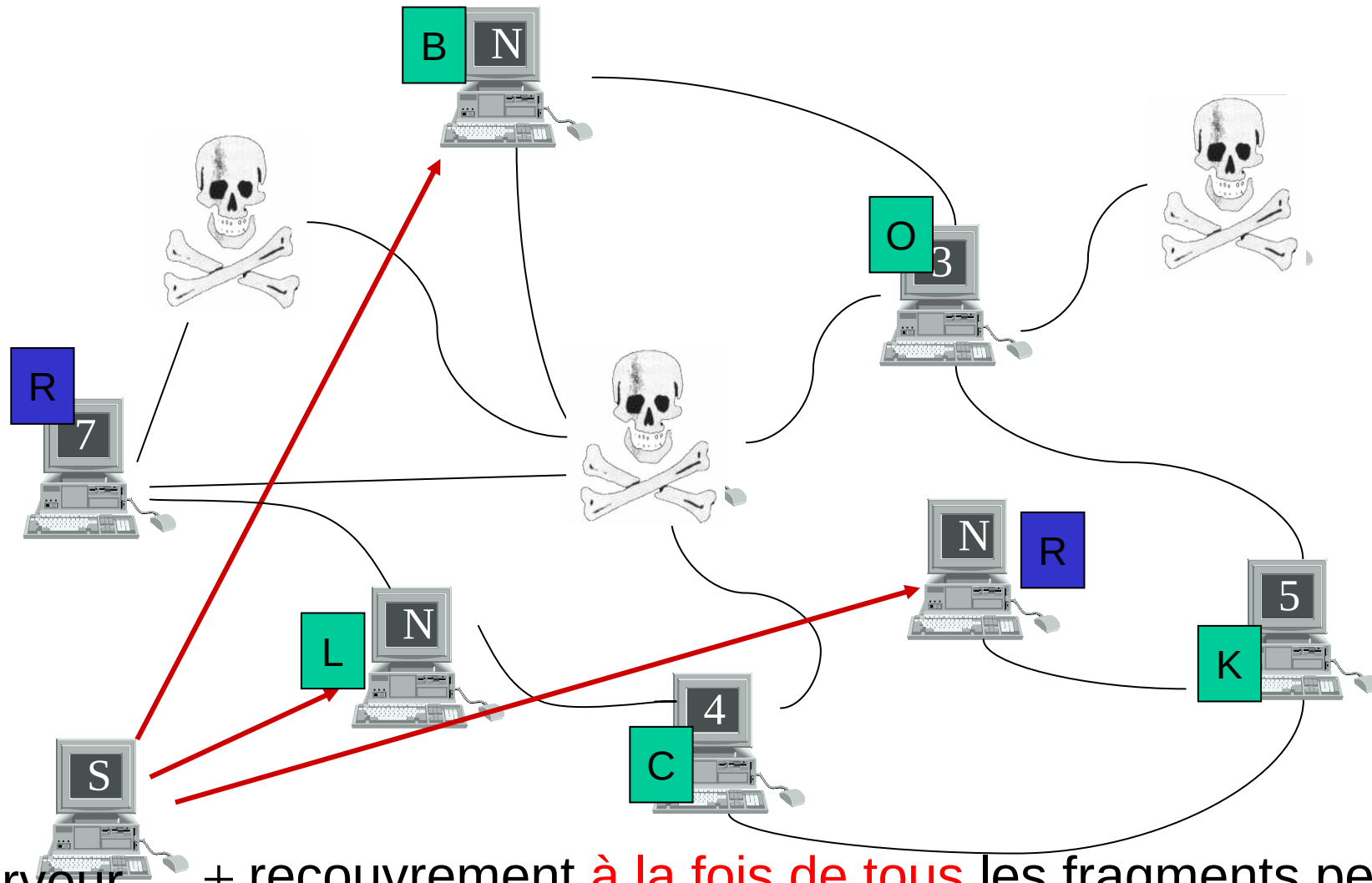


1- Mécanisme de recouvrement centralisé

Téléchargement des s fragments nécessaires



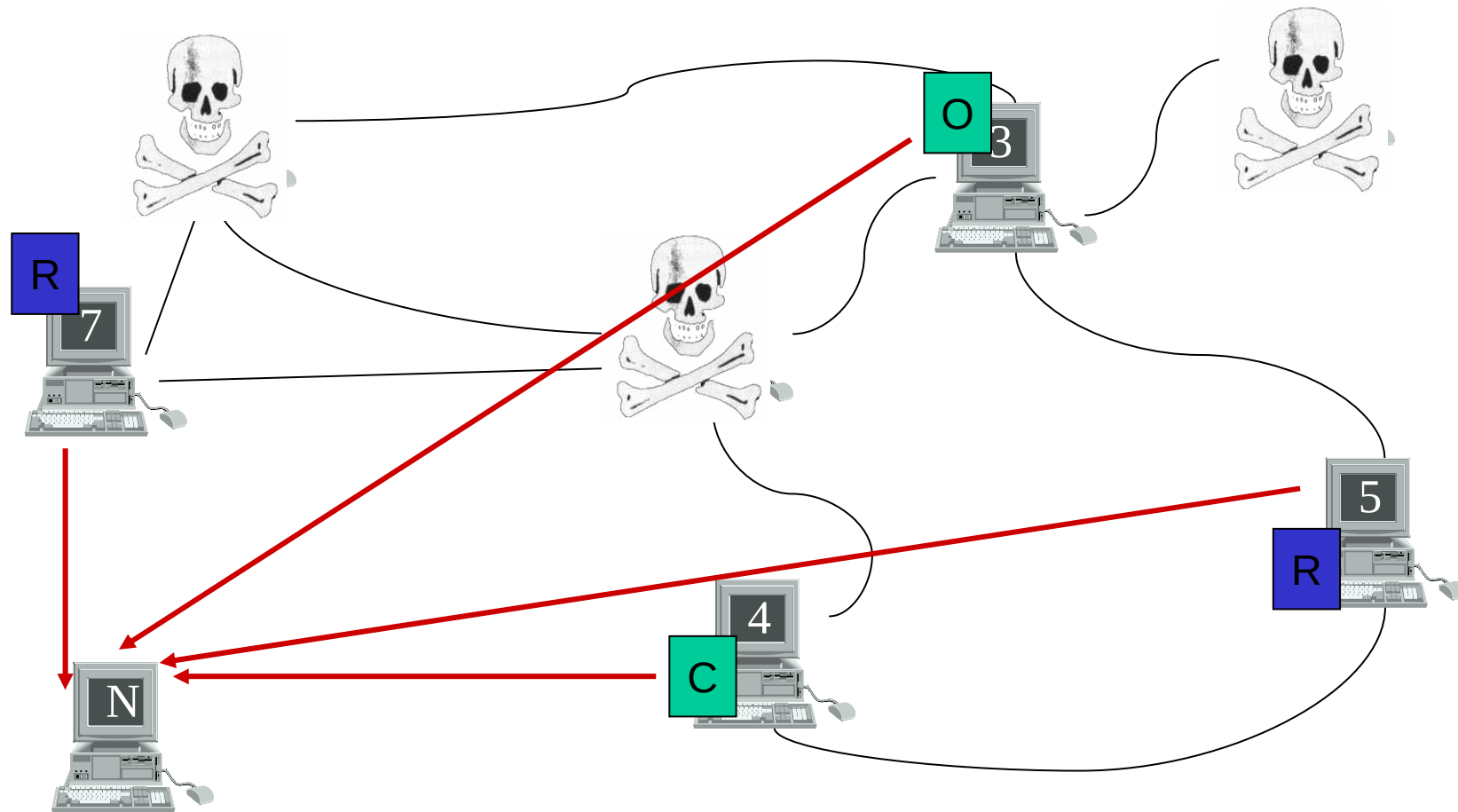
1- Mécanisme de recouvrement centralisé



Serveur + recouvrement **à la fois de tous** les fragments perdus
centralisé – **besoin** d'une autorité centrale

1- Mécanisme de recouvrement distribué

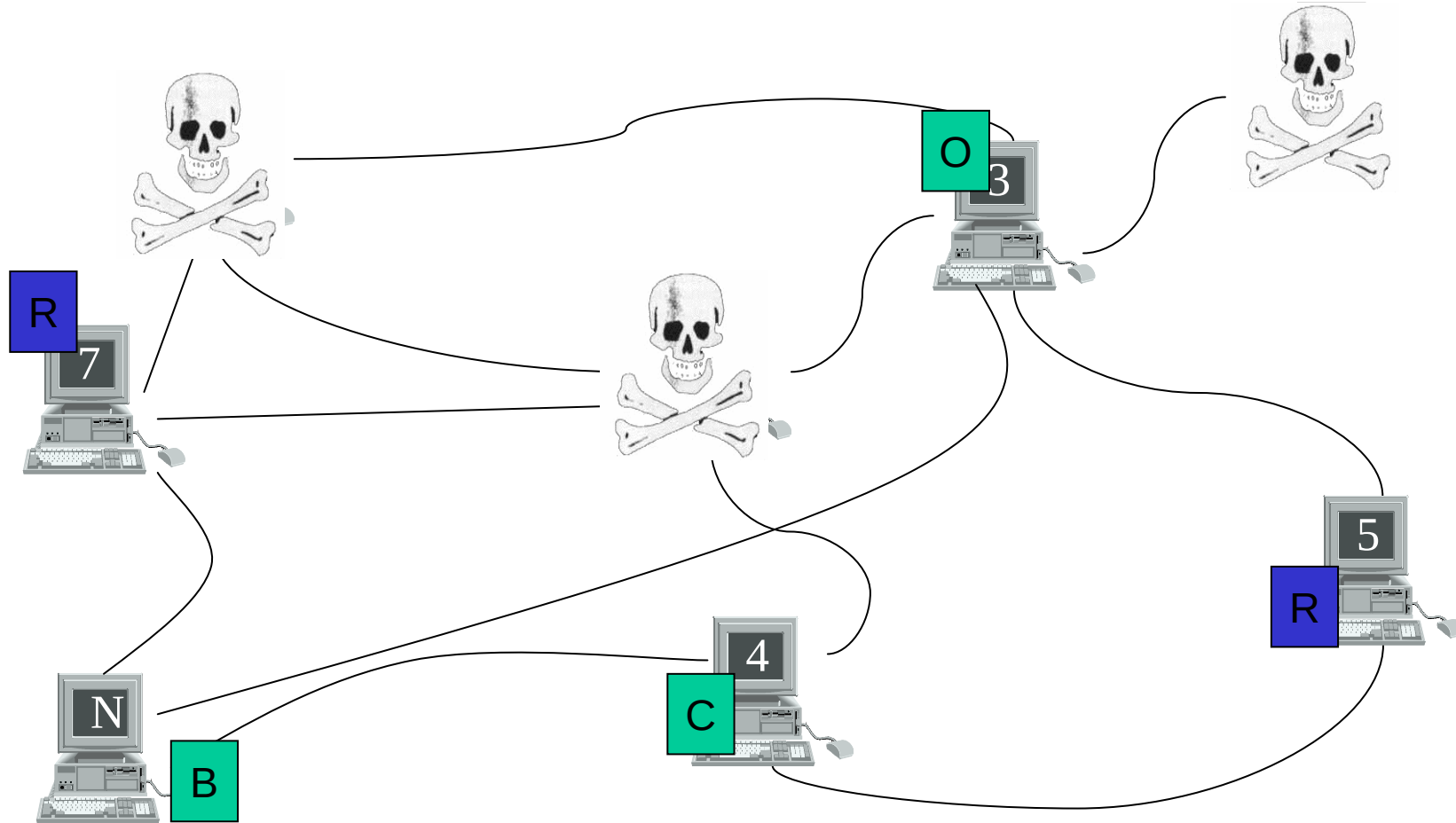
Téléchargement des s fragments nécessaires



Nouveau pair

1- Mécanisme de recouvrement distribué

- + pas besoin d'un serveur centralisé
- reconstruction d'un seul fragment au lieu de tous les fragments



Nouveau pair

Plan

- Motivation & Introduction
- Contributions
- Description du système et hypothèses
- Modèles Markoviens pour
 - ▶ mécanisme de recouvrement centralisé
 - ▶ mécanisme de recouvrement distribué
- Résultats principaux
- Conclusion

Contributions

- Évaluation des systèmes de stockage en termes de disponibilité et de pérennité des données en utilisant une chaîne de Markov absorbante
 - ▶ Gestion centralisée/distribuée de la reconstruction
 - ▶ Distributions réalistes des durées de connexions/déconnexions
- Paramétrage du système pour la satisfaction de critères de fiabilité et de disponibilité
 - ▶ Niveau de redondance r
 - ▶ Seuil k de déclenchement de la reconstruction des données
- Impact du paramètre s

Description du système et hypothèses

- Nombre de pairs suffisamment grand
- Fragments d'un même bloc répartis sur des pairs indépendants
- Redondance (Reed-Solomon, Tornado, ...)
- Politique de recouvrement
 - ▶ Quand déclencher la réparation?

Description du système et hypothèses

- Deux politiques de recouvrement
 - ▶ Réparation immédiate des données perdues ($k=1$)
 - + fiabilité ↗ – bande passante ↗
 - ▶ Réparation retardée jusqu'à ce que la perte atteigne un seuil ($1 < k \leq r$)
 - + bande passante ↘ – besoin de plus de redondance, donc plus d'espace de stockage
- Durées de déconnexions
 - v.a. i.i.d., distribution exponentielle de paramètre λ

cette hypothèse est en accord avec l'analyse de la disponibilité des machines en : S. Ramabhadran, J. Pasquale, Analysis of long-running replicated systems, in: Proc. of IEEE Infocom '06, Spain. Etude de la trace *All-pairs-ping* pour les pairs de Planet-lab.

Description du système et hypothèses

■ Durées de connexions

- ▶ v.a. i.i.d., distribution hyper-exponentielle de n phases
cette hypothèse est en accord avec l'analyse de la disponibilité des machines effectuée sur trois traces de données : D. Nurmi, J. Brevik, R. Wolski, Modeling machine availability in enterprise and wide-area distributed computing environments, Tech. Rep. CS2003, California Santa Barbara

■ Durées de réparation

- ▶ v.a. i.i.d., distribution exponentielle de paramètre β

■ La probabilité que des pairs stockant encore leurs données se reconnectent est p

Modélisation du système de stockage

- L'état du système est représenté par le **nombre de fragments disponibles** d'un document
- Sous les hypothèses considérées, l'état du système représenté est une chaîne de Markov absorbante
- ***Etat Absorbant*** → moins de s fragments sont disponibles → bloc est perdu !
- Mesures de performances étudiées
 - ▶ **Durée de vie** d'un document (**distribution, moyenne**)

$$T_c^h(\mathcal{E}_I) := \inf\{t \geq 0 : X_c^h(t) = a | X_c^h(0) \in \mathcal{E}_I\}$$

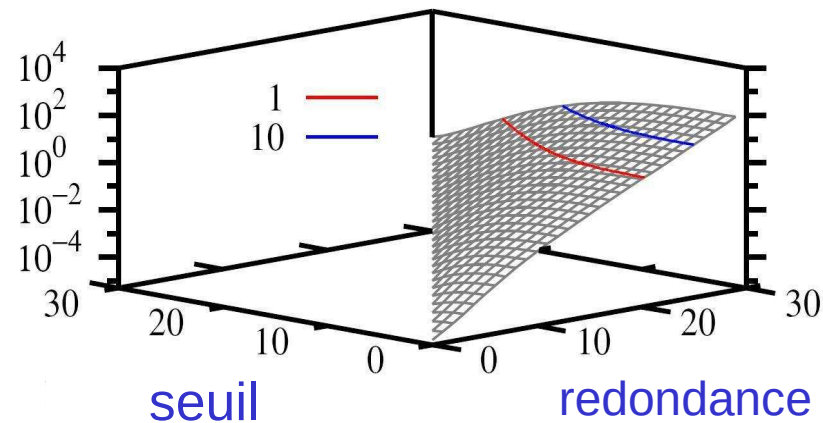
Modélisation du système de stockage

- La disponibilité des données est caractérisée par deux paramètres
 - ▶ le nombre moyen de fragments disponibles
 - ▶ la fraction de la durée de vie du bloc où au moins **m** fragments redondants sont disponibles

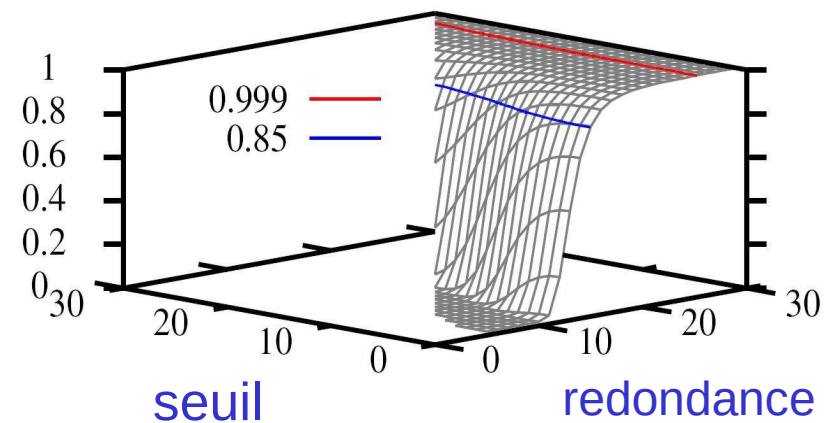
$$M_{c,1}^h(\mathcal{E}_I) := \sum_{J=s}^{s+r} J \frac{\mathbb{E} [T_c^h(\mathcal{E}_I, \mathcal{E}_J)]}{\mathbb{E} [T_c^h(\mathcal{E}_I)]}, \quad M_{c,2}^h(\mathcal{E}_I) := \sum_{J=m}^{s+r} \frac{\mathbb{E} [T_c^h(\mathcal{E}_I, \mathcal{E}_J)]}{\mathbb{E} [T_c^h(\mathcal{E}_I)]}.$$

Evaluation (gestion centralisée)

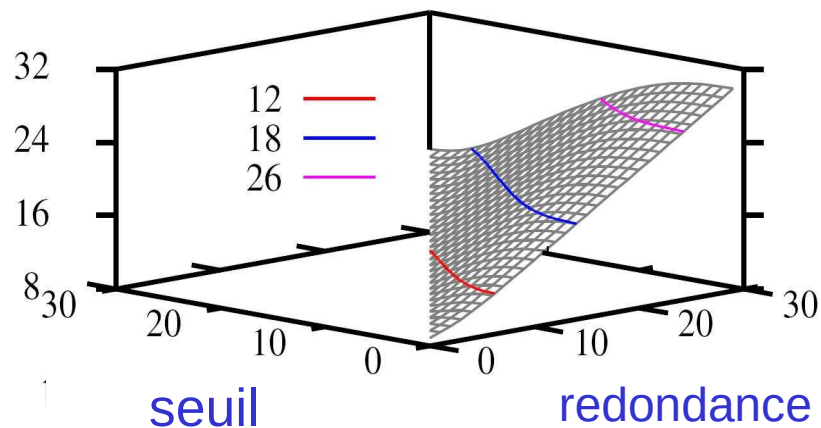
Durée de vie moyenne



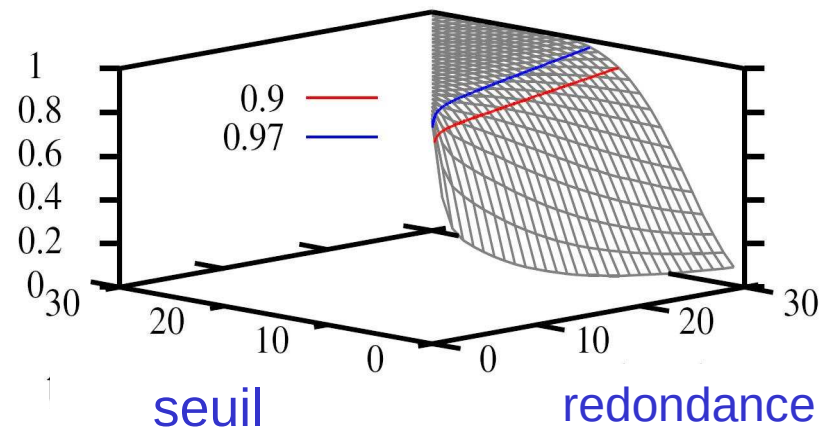
Pérennité au bout d'un an



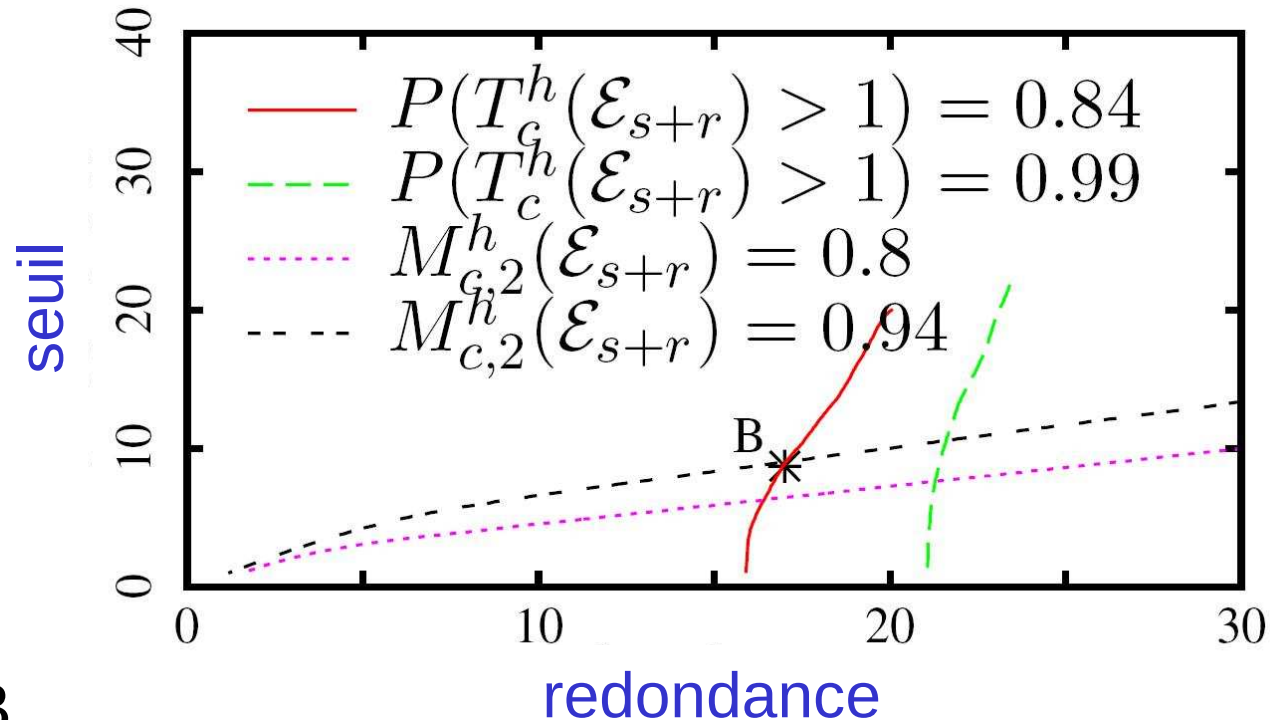
Disponibilité moyenne



Disponibilité (%) au dessus du seuil



Exemple de Paramétrage

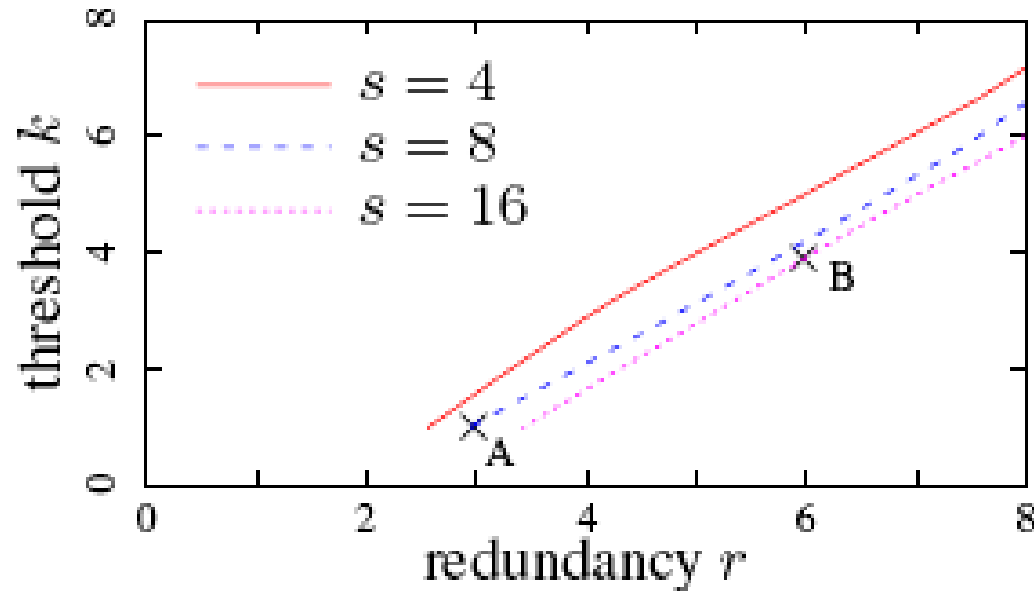


■ Point B

- ▶ à chaque $s = 8$ fragments s'ajoutent $r = 17$ fragments redondants
- ▶ seuil de reconstruction égal à $k = 9$

▶ 84% des documents subsistent au bout d'un an
94% des documents ont au moins 16 fragments disponibles

Impact du paramètre s



(a) LMG data, $P(T_d^h(\mathcal{E}_{s+r}) > 10) = 0.99$

- **A** ($s = 8, r = 3, k = 1$) et **B** ($s = 16, r = 6, k = 4$)
 - ▶ Même volume de stockage → même overhead = 0.375
 - ▶ Mais seuil avec A inférieur au seuil avec B
 - bande passante utilisée avec A supérieure à celle utilisée avec B, bien que A et B offrent la même garantie de vie et le même volume de stockage

Conclusions & Extensions

- Évaluation des systèmes de stockage en termes de disponibilité et de pérennité des données en utilisant une chaîne de Markov absorbante
 - ▶ Gestion centralisée/distribuée de la reconstruction
 - ▶ Distributions réalistes des durées de connexions/déconnexions
- Paramétrage du système pour la satisfaction de critères de fiabilité et de disponibilité
- Impact du paramètre s
- Extensions
 - ▶ Généralisation de la distribution du temps de réparation (hypo-exponentielle ou Erlang-2 généralisée)
 - ▶ Implémentation du protocole avec Ns-2.33

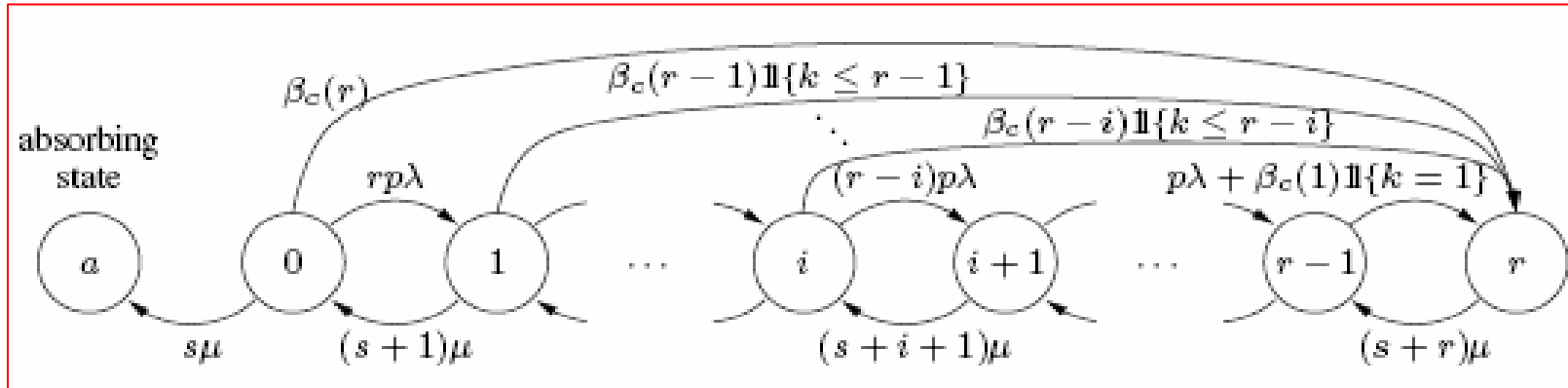


Merci pour votre attention !

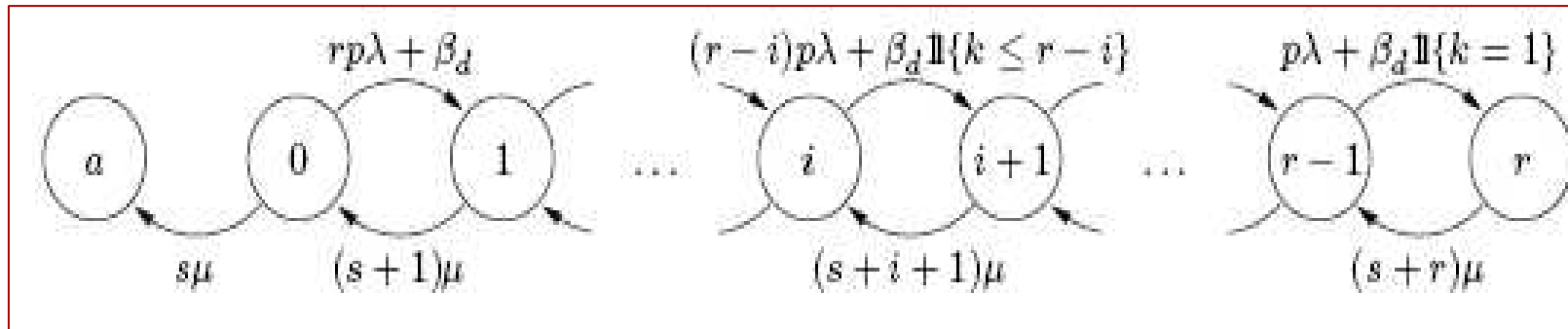
Questions?

Simple Examples of the AMC

1- Centralized Recovery System: $n=1$, on-times exp. dist. with μ



2- Distributed Recovery System: $n=1$, on-times exp. dist. with μ



P2P Storage Systems

- Examples of large-scale systems:

- ▶ Intermemory
- ▶ Freenet
- ▶ Oceanstore (Berkeley)
- ▶ Farsite (Microsoft)
- ▶ TotalRecall

- Commercial systems:

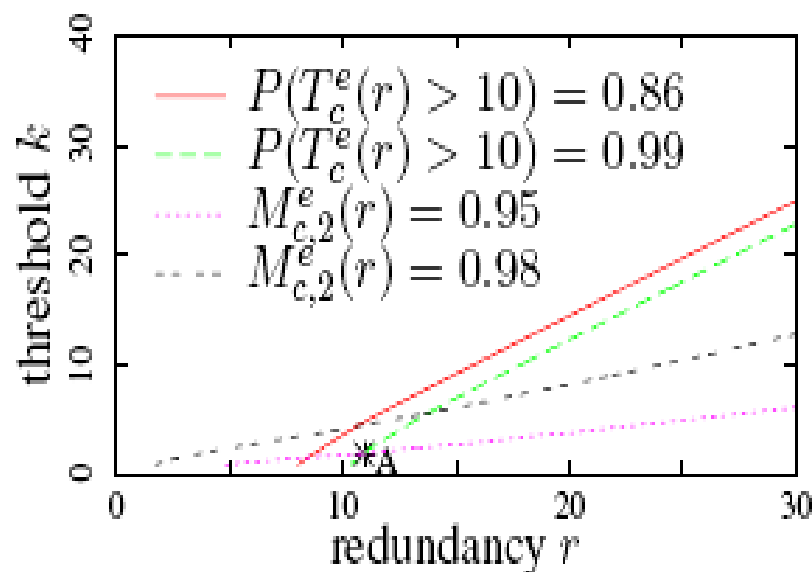
- ▶ UbiStorage[©] (www.ubistorage.com)

Table 2

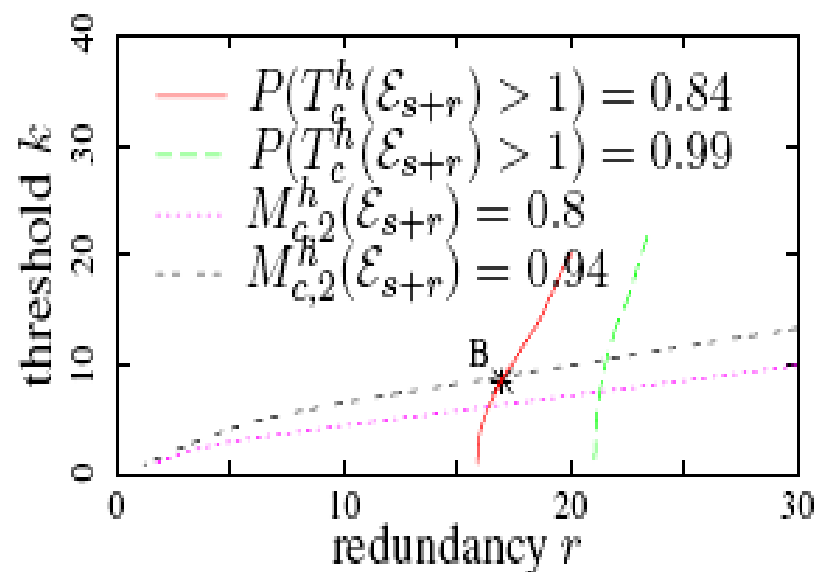
Data sets characteristics and corresponding peers parameters values

Data set	<i>LMG</i>	<i>CSIL</i>	<i>Condor</i>	<i>All-pairs-ping</i>
Context	Internet	LAN	Internet	PlanetLab
Covered period	3 months	8 weeks	6 weeks	21 months
Number of peers	1170	83	210	200–550
On-times distribution	H_3 [19]	H_3 [19]	H_2 [19]	Exp. [15]
	(best fit)	(best fit)	(best fit)	(reasonable)
On-times parameters				
p_1	0.282	0.464	0.592	1
p_2	0.271	0.197	0.408	–
p_3	0.447	0.339	–	–
$1/\mu_1$ (hours)	910.7	250.3	0.094	181
$1/\mu_2$ (hours)	0.224	1.425	3.704	–
$1/\mu_3$ (hours)	199.8	33.39	–	–
Mean on-time (hours)	352.2	127.7	1.567	181 [15]
Mean off-time (hours)	48.43 [14]	48	1.567 or 0.522	61 [15]
Percentage of on-times	0.879	0.727	0.5 or 0.75	0.750
Persistence probability p	0.4	0.4	0.8	0.4

Engineering the system



(a) All-pairs-ping data



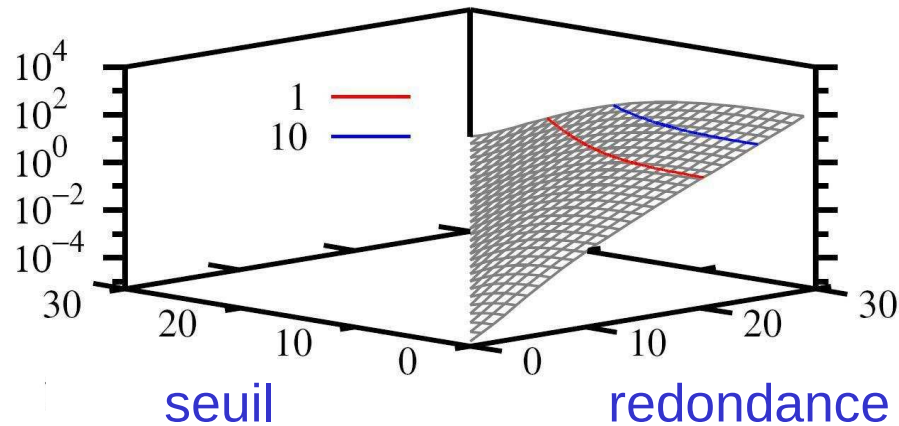
(b) Condor data, $1/\lambda = 0.522$

■ Point B

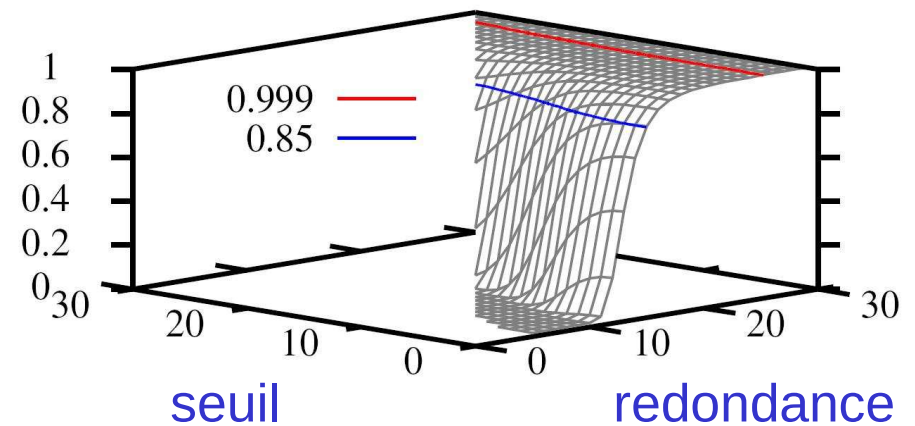
- ▶ $s = 8$ fragments , we add $r = 17$ redundant fragments
- ▶ Cost: storage overhead = $17/8 = 2.125$
- ▶ Thresold of recovery $k = 9$
- ▶ the document still available more than one year with Prob. = 84%
and during 94% of its lifetime, more than $r - k = 8$ redundant fragments are available

Evaluation (gestion centralisée)

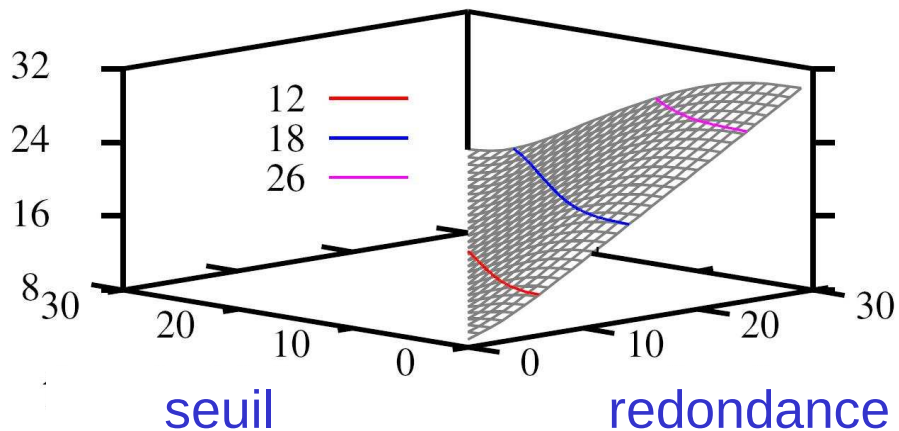
Durée de vie moyenne



Pérennité au bout d'un an



Disponibilité moyenne



Disponibilité (%)
au dessus du seuil

