HAL
open science

Performance Analysis of Centralized versus Distributed Recovery Schemes in P2P Storage Systems

Abdulhalim Dandoush, Sara Alouf, Philippe Nain

To cite this version:

Abdulhalim Dandoush, Sara Alouf, Philippe Nain. Performance Analysis of Centralized versus Distributed Recovery Schemes in P2P Storage Systems. [Research Report] RR-6769, INRIA. 2008, pp.14. inria-00346503

HAL Id: inria-00346503

https://inria.hal.science/inria-00346503

Submitted on 11 Dec 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Performance Analysis of Centralized versus Distributed Recovery Schemes in P2P Storage Systems

Abdulhalim Dandoush — Sara Alouf — Philippe Nain

N° 6769

December 2008

Thème COM

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6769--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Performance Analysis of Centralized versus Distributed Recovery Schemes in P2P Storage Systems

Abdulhalim Dandoush , Sara Alouf , Philippe Nain

Thème COM — Systèmes communicants
Équipe-Projet MAESTRO

Rapport de recherche n° 6769 — December 2008 — 14 pages

Abstract: This report studies the performance of Peer-to-Peer Storage Systems (P2PSS) in terms of data lifetime and availability. Two schemes for recovering lost data are modeled through absorbing Markov chains and their performance are evaluated and compared. The first scheme relies on a centralized controller that can recover multiple losses at once, whereas the second scheme is distributed and recovers one loss at a time. The impact of each system parameter on the performance is evaluated, and guidelines are derived on how to engineer the system and tune its key parameters in order to provide desired lifetime and/or availability of data. We find that, in stable environments such as local area or research laboratory networks where machines are usually highly available, the distributed-repair scheme offers a reliable, scalable and cheap storage/backup solution. This is in contrast with the case of highly dynamic environments, where the distributed-repair scheme is inefficient as long as the storage overhead is kept reasonable. P2PSS with centralized-repair scheme are efficient in any environment but have the disadvantage of relying on a centralized authority. Our analysis also suggests that the use of large size fragments reduces the efficiency of the recovery mechanism.

Key-words: peer-to-peer storage systems, recovery process, absorbing continuous-time Markov chain, performance evaluation

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Analyse de performance de mécanismes centralisé et distribué de récupération de données dans des systèmes pair-à-pair dédiés au stockage

Résumé : Ce rapport étudie les performances des systèmes pair-à-pair dédiés au stockage (archivage ou sauvegarde) en termes de durée de vie des données et de leur disponibilité. Deux mécanismes de récupération des données perdues sont modélisés par une chaîne de Markov absorbante et leurs performances sont évaluées et comparées. Le premier mécanisme nécéssite l'utilisation d'un serveur, pouvant ainsi récupérer plusieurs données à la fois, alors que le second mécanisme est distribué mais ne récupère qu'une seule donnée perdue à la fois. Nous évaluons l'impact de chaque paramètre sur la performance du système et montrons comment nos résultats peuvent être utilisés de sorte à garantir que la qualité de service pré-requise soit pourvue. Nous constatons que, dans des environnements stables, tels que les réseaux locaux ou ceux dédiés à la recherche, où les machines sont généralement hautement disponibles, la réparation distribuée offre une solution de sauvegarde fiable, performante et peu coûteuse. Toutefois, quand la dynamique des pairs est prononcée, la réparation distribuée devient peu efficace, surtout à des niveaux de redondance raisonnables. Quant au mécanisme centralisé de réparation des données, il s'est avéré efficace dans n'importe quel environnement, l'inconvénient étant de reposer sur l'utilisation d'un serveur centralisé, avec ce que ceci implique de gestion et de sensibilité aux attaques et défailances. Notre étude suggère aussi qu'une fragmentation grossière des données réduirait l'efficacité du mécanisme de récupération.

Mots-clés : systèmes de stockage pair-à-pair, réparation des données, chaîne de Markov absorbante en temps continu, évaluation de performance

Contents

1 Introduction **3**

2 Related Work and Background **4**

3 System Description, Assumptions and Notation **4**

4 Centralized Repair Systems **6**
4.1 Data Lifetime 8
4.2 Data Availability 9

5 Distributed Repair Systems **9**

6 Numerical Results **10**
6.1 Validation of (4) 11
6.2 Performance analysis 11
6.3 Setting the system's key parameters 12
6.4 Impact of the size of fragments 13

7 Conclusion **13**

1 Introduction

The growth of storage volume, bandwidth, and computational resources for PCs has fundamentally changed the way applications are constructed, and has inspired a new class of storage systems that use distributed peer-to-peer (P2P) infrastructures. Although scalable and economically attractive compared to traditional systems, these storage systems pose many problems such as reliability, confidentiality and availability. To ensure data reliability and availability in such dynamic systems, redundant data is inserted in the system. However, using redundancy mechanisms without repairing lost data is not efficient, as the level of redundancy decreases when peers leave the system. Consequently, P2P storage systems need to compensate the loss of data by continuously storing additional redundant data onto new hosts. Systems may rely on a central authority that reconstructs fragments when necessary; these systems will be referred to as *centralized-recovery systems*. Alternatively, secure agents running on new hosts can reconstruct by themselves the data to be stored on the hosts disks. Such systems will be referred to as *distributed-recovery systems*. Regardless of the recovery mechanism used, two repair policies can be adopted *eager* and *lazy* policy. The description of these recovery schemes and their policies are presented in Section 3.

The aim of this report is to develop and evaluate mathematical models to characterize fundamental performance metrics (data lifetime and availability) of P2P storage systems. Our contributions are as follows

- Analysis of centralized and distributed recovery mechanisms.
- Proposition of a general model that captures the behavior of both eager and lazy repair policies and both replication-based and erasure code-based systems, and accommodates both temporary and permanent disconnections of peers.
- Numerical investigation using realistic parameters values.

- Guidelines on how to engineer the P2PSS in order to satisfy given requirements.

In the following, Section 2 briefly reviews related work and Section 3 introduces the notation and assumptions used throughout the report. Sections 4 and 5 are dedicated to the modeling of the centralized- and distributed-recovery mechanism, respectively. In Section 6, we provide some numerical results showing the performance of the centralized and decentralized schemes. Section 7 concludes the report.

2 Related Work and Background

The literature on the architecture and file system of distributed storage systems is abundant (see [11, 3]; non-exhaustive list), but to the best of our knowledge, there has been no prior work on the modeling of the recovery process in P2PSS. A few studies have developed analytical models with the goal of understanding the trade-offs between the availability and lifetime of the files and the redundancy involved in storing the data. From these we cite [9] whose main purpose of [9] is the analysis of a storage system using replication for data reliability. The work [9] is the closest to ours even though the model and analysis developed therein do not apply for erasure-coded systems (we will see later that our models apply to either replicated or erasure-coded systems).

The authors of [9] develop a Markov chain analysis, then derive an expression for the lifetime of the replicated state and study the impact of bandwidth and storage limits on the system. However – and these are major differences with the work presented here, transient disconnections are not considered in their model, the recovery process is considered to be exponentially distributed for the aim of simplification, and only the distributed-repair scheme is considered. Another contribution of [9] is the analysis of the *All-pairs-ping* data set [12] that reports measures of both uptime and downtime for PlanetLab [8] nodes. The authors found that an exponential distribution is a reasonable fit for both uptime and downtime. This conjecture comes to support one of the key assumptions of the model presented in our report, namely that "node participation can be modeled by an exponential distribution".

In our previous work [1], simplifying assumptions have been considered while modeling the P2PSS, mainly that the recovery process is exponentially distributed. This assumption is relaxed in this report, and it will be shown later on that a more precise, more realistic modeling is possible.

3 System Description, Assumptions and Notation

We consider a distributed *storage* system which peers randomly join and leave. The following assumptions on the P2PSS design will be enforced throughout the report:

- A block of data D is partitioned into s equally sized fragments to which, using erasure codes (e.g. [10]), r redundant fragments are added. The case of replication-based redundancy is equally captured by this notation, after setting $s = 1$ and letting the r redundant fragments be simple replicas of the unique fragment of the block. This notation – and hence our modeling – is general enough to study both replication-based and erasure code-based storage systems.

- Mainly for privacy issues, a peer can store at most one fragment of any data D.

- We assume the system has perfect knowledge of the location of fragments at any given time, e.g. by using a Distributed Hash Table (DHT) or a central authority.
- The system keeps track of only the latest known location of each fragment.
- Over time, a peer can be either *connected* to or *disconnected* from the storage system. At reconnection, a peer may or may not still store its fragments. We denote by p the probability that a peer that reconnects still stores its fragments.
- The number of connected peers at any time is typically much larger than the number of fragments associated with D, i.e., $s + r$. Therefore, we assume that there are always at least $s + r$ connected peers – hereafter referred to as *new* peers – which are ready to receive and store fragments of D.

We refer to as *on-time* (resp. *off-time*) a time-interval during which a peer is always connected (resp. disconnected). During a peer's off-time, the fragments stored on this peer are momentarily unavailable to the users of the storage system. At reconnection, and according to the assumptions above, the fragments stored on this peer will be available only with probability p (and with probability $1 - p$ it is lost). In order to improve data availability and increase the reliability of the storage system, it is therefore crucial to recover from losses by continuously monitoring the system and adding redundancy whenever needed.

We will investigate the performance of two different repair policies: the *eager* and the *lazy* repair policies. In the eager policy, a fragment of D is reconstructed as soon as one fragment has become unavailable due to a peer disconnection. In the lazy policy, the repair is delayed until the number of unavailable fragments reaches a given threshold, denoted k. In the latter case, we must have $k \leq r$ since D is lost if more than r fragments are missing from the storage system. Both repair policies can be represented by the threshold parameter $k \in \{1, 2, \ldots, r\}$, where k can take any value in the set $\{2, \ldots, r\}$ in the lazy policy and $k = 1$ in the eager policy. Any repair policy can be implemented either in a centralized or a distributed way. In the following description, we assume that the system misses k fragments so that lost fragments have to be restored.

In the centralized implementation, a central authority will: (1) download *in parallel* s fragments from the peers which are connected, (2) reconstruct at once all the unavailable fragments, and (3) upload them all *in parallel* onto as many new peers for storage. Step 2 executes in a negligible time compared to the execution time of Steps 1 and 3 and will henceforth be ignored in the modeling. Step 1 (resp. Step 3) execution completes when the last fragment completes being downloaded (resp. uploaded).

In the distributed implementation, a secure agent on one new peer is notified of the identity of *one* out of the k unavailable fragments for it to reconstruct it. Upon notification, the secure agent (1) downloads s fragments of D from the peers which are connected to the storage system, (2) reconstructs the specified fragment and stores it on the peer's disk; (3) the secure agent then discards the s downloaded fragments so as to meet the privacy constraint that only one fragment of a block of data is held by a peer. This operation iterates until less than k fragments are sensed unavailable and stops if the number of missing fragments reaches $k - 1$. The recovery of one fragment lasts mainly for the execution time of Step 1; the recovery is completed then as soon as the last fragment (out of s) completes being downloaded.

In both implementations, once a fragment is reconstructed, any other copy of it that "reappears" in the system due to a peer reconnection is simply ignored, as only

one location (the newest) of the fragment is recorded in the system. Similarly, if a fragment is unavailable, the system knows of only one disconnected peer that stores the unavailable fragment.

Given the system description, data D can be either *available*, *unavailable* or *lost*. Data D is said to be available if any s fragments out of the $s + r$ fragments can be downloaded by the users of the P2PSS. Data D is said to be unavailable if *less than* s fragments are available for download, however the missing fragments to complete D are located at a peer or a central authority on which a recovery process is ongoing. Data D is said to be lost if there are *less than* s fragments in the system including the fragments involved in a recovery process. We assume that, at time $t = 0$, at least s fragments are available so that the document is initially available.

We now introduce the assumptions considered in our models.

Assumption 1: We assume that the successive durations of on-times (resp. off-times) of a peer are independent and identically distributed (iid) random variables (rvs) with a common exponential distribution function with parameter $\mu > 0$ (resp. $\lambda > 0$); this assumption is in agreement with the analysis in [9]. We further assume that peers behave independently of each other.

Assumption 2: We assume that the successive download (resp. upload) durations of a fragment are iid rvs with a common exponential distribution function with parameter α (resp. β). Fragments downloads/uploads are not correlated.

A consequence of Assumption 2 is that each of the durations of the centralized and the distributed recovery processes is a rv following a hypo-exponential distribution [6]. Indeed, each of these durations is the summation of independently distributed exponential rvs ($s + k$ in the centralized scheme if k fragments are to be reconstructed, and s in the distributed scheme) having each its own rate. This is a fundamental difference with [9, 1] where the recovery process is assumed to be an exponential distribution. We believe that Assumption 2 is fairly realistic and that the subsequent modeling is much closer to reality than what was done in [9, 1].

We conclude this section by a word on the notation: a subscript "c" (resp. "d") will indicate that we are considering the centralized (resp. distributed) scheme. The notation $\boldsymbol{e}_j^i$ refers to a *row* vector of dimension j whose entries are null except the i-th entry that is equal to 1; the notation $\mathbf{1}_j$ refers to a *column* vector of dimension j whose each entry is equal to 1. Last, $\mathbb{1}\{A\}$ is the characteristic function of event A.

4 Centralized Repair Systems

We will focus on a single block of data D, and pay only attention to peers storing fragments of this block.

Let $X_c(t)$ and $Y_c(t)$ be two rvs denoting respectively the number of fragments in the system that are available for download and the state of the recovery process. Recall that, when k fragments are to be reconstructed, the recovery process consists of a series of $s+k$ exponential distributions that can be seen as $s+k$ stages. We denote $Y_c(t) = j$ ($j = 0, 1, \ldots, k - 1$) to express that j exponential rvs have been realized at time t, so that $s + k - j$ are still to go. When the last stage is completed, the recovery process is completed and $Y_c(t) = 0$. Given that there could be as much as $s + r$ fragments to be reconstructed, the process $Y_c(t)$ takes value in the set $\{0, 1, \ldots, 2s + r - 1\}$. As for $X_c(t)$, it takes value in the set $\{0, 1, \ldots, s + r\}$.

Consider now the joint process $(X_c(t), Y_c(t))$. When $X_c(t) \geq s$, data D is *available*, regardless of $Y_c(t)$. When $X_c(t) < s$ but $X_c(t) + Y_c(t) \geq s$, D is *unavailable*. When $X_c(t) + Y_c(t) < s$, D is *lost*. The latter situation will be modeled by a single state a. Introduce the set

$$\begin{aligned}
\mathcal{T}_c := \{ \quad & \left.\begin{array}{l} (0,s),(0,s+1),\ldots,(0,2s+r-1), \\ (1,s-1),(1,s),\ldots,(1,2s+r-2), \\ \ldots, \end{array}\right\} \quad D \text{ is unavailable} \\
& \left.\begin{array}{l} (s,0),(s,1),\ldots,(s,s+r-1), \\ (s+1,0),(s+1,1),\ldots,(s+1,s+r-2), \\ \ldots,(s+r-1,0),(s+r-1,1),\ldots,(s+r-1,s), \\ (s+r,0)\ \} \end{array}\right\} \quad D \text{ is available} \\
|\mathcal{T}_c| = (s+r)^2 - r(r-1)/2 + 1.
\end{aligned}$$

Thanks to the assumptions made in Sect. 3, it is easily seen that the two-dimensional process $\{(X_c(t), Y_c(t)),\ t \geq 0\}$ is an absorbing homogeneous Continuous-Time Markov Chain (CTMC) with transient states the elements of $\mathcal{T}_c$ and with a single absorbing state a representing the situation when D is lost. Without loss of generality, we assume that $X_c(0) \geq s$. The infinitesimal generator has the following canonical form

$$\begin{array}{cc} & \begin{array}{cc} \mathcal{T}_c & a \end{array} \\ \begin{array}{c} \mathcal{T}_c \\ a \end{array} & \left(\begin{array}{c|c} \boldsymbol{Q}_c & \boldsymbol{R}_c \\ \hline \boldsymbol{0} & 0 \end{array}\right) \end{array}$$

where $\boldsymbol{R}_c$ is a non-zero column vector of size $|\mathcal{T}_c|$, and $\boldsymbol{Q}_c$ is $|\mathcal{T}_c|$-by-$|\mathcal{T}_c|$ matrix. The elements of $\boldsymbol{R}_c$ are the transition rates between the transient states $(i,j) \in \mathcal{T}_c$ and the absorbing state a, namely, $r_c(i,j) = (s-j)\mu$, for $i = 1,\ldots,s$, and $j = s-i,\ldots,s-1$. The elements of $\boldsymbol{R}_c$ are lexicographically ordered alike the order in $\mathcal{T}_c$. The diagonal elements of $\boldsymbol{Q}_c$ are each the total transition rate out of the corresponding transient state. The other elements of $\boldsymbol{Q}_c$ are the transition rates between each pair of transient states. The non-zero elements of $\boldsymbol{Q}_c$ are:

$$q_c((i,j),(i-1,j)) = \begin{cases} i\mu, & \text{for} \quad i = 1,\ldots,s, j = s,\ldots,2s+r-1-i; \\ & \text{or} \quad i = s+1,\ldots,s+r-1, \\ & \qquad j = 0,\ldots,2s+r-1-i; \\ & \text{or} \quad i = s+r, j = 0; \\ (i+j-s)\mu, & \text{for} \quad i = 2,\ldots,s, j = s+1-i,\ldots,s-1. \end{cases}$$

$$q_c((i,j),(i,j+1)) = \begin{cases} (s-j)\alpha, & \text{for} \quad i = s,\ldots,s+r-k, j = 0; \\ & \text{or} \quad i = 1,\ldots,s-1, j = s-i,\ldots,s-1; \\ & \text{or} \quad i = s,\ldots,s+r-1, j = 1,\ldots,s-1; \\ (2s+r-i-j)\beta, & \text{for} \quad i = 0,\ldots,s+r-2, \\ & \qquad j = s,\ldots,2s+r-2-i. \end{cases}$$

$$\begin{aligned}
& q_c((i,2s+r-1-i),(s+r,0)) = \beta, && \text{for } i = 0,\ldots,s+r-1. \\
& q_c((i,j),(i+1,j)) = (s+r-i)\lambda p, && \text{for } i = 1,\ldots,s, j = s-i,\ldots,s-1; \\
& && \text{or } i = s+1,\ldots,s+r-2, j = 0,\ldots,s-1. \\
& q_c((s+r-1,j),(s+r,0)) = \lambda p, && \text{for } j = 0,\ldots,s-1. \\
& q_c((i,j),(i,j)) = -r_c(i,j) - \textstyle\sum_{(i',j')\in\mathcal{T}_c-\{(i,j)\}} q_c((i,j),(i',j')), && \text{for } (i,j) \in \mathcal{T}_c.
\end{aligned}$$

Note that $\boldsymbol{Q}_c$ is not an infinitesimal generator since entries in some rows do not sum up to 0. For illustration purposes, we depict in Fig. 1 an example of the absorbing CTMC with its non-zero transition rates when $s = 2$, $r = 2$, and $k = 2$.

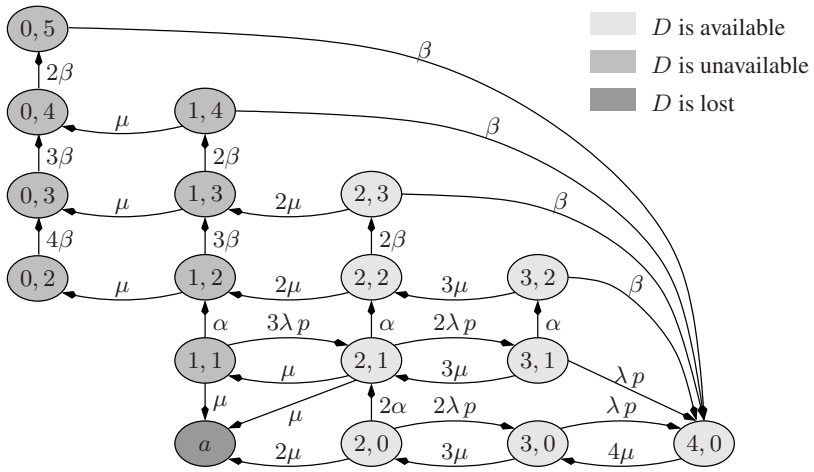

Figure 1: The Markov chain $\{(X_c(t), Y_c(t)),\ t \geq 0\}$ when $s = 2$, $r = 2$, $k = 2$.

4.1 Data Lifetime

This section is devoted to the analysis of the lifetime of D. Let $T_c(i,j) := \inf\{t > 0 : (X_c(t), Y_c(t)) = a | (X_c(0), Y_c(0)) = (i,j)\}$ be the time until absorption in state a, or equivalently the time until D is lost, given that the initial state of D is (i,j). In the following, $T_c(i,j)$ will be referred to as the *conditional block lifetime*. We are interested in $P(T_c(i,j) \leq x)$ and $\mathrm{E}[T_c(i,j)]$, respectively the probability distribution of the conditional block lifetime and its expectation, given that $(X_c(0), Y_c(0)) = (i,j) \in \mathcal{T}_c$. From the theory of absorbing Markov chains, we know that (e.g. [7, Lemma 2.2])

$$P(T_c(i,j) \leq x) = 1 - \boldsymbol{e}_{|\mathcal{T}_c|}^{\mathrm{ind}(i,j)} \cdot \exp\left(x\boldsymbol{Q}_c\right) \cdot \mathbf{1}_{|\mathcal{T}_c|}, \quad x > 0,\ (i,j) \in \mathcal{T}_c \tag{1}$$

where $\mathrm{ind}(i,j)$ refers to the index of the state $(i,j) \in \mathcal{T}_c$ in the matrix $\boldsymbol{Q}_c$. Recall that the elements of $\boldsymbol{Q}_c$ are numbered according to the lexicographic order. Definitions of vectors $\boldsymbol{e}_i^j$ and $\mathbf{1}_i$ are given at the end of Section 3. Observe that the term $\boldsymbol{e}_{|\mathcal{T}_c|}^{\mathrm{ind}(i,j)} \cdot \exp\left(x\boldsymbol{Q}_c\right) \cdot \mathbf{1}_{|\mathcal{T}_c|}$ in the r.h.s. of (1) is nothing but the summation of all $|\mathcal{T}_c|$ elements in row $\mathrm{ind}(i,j)$ of matrix $\exp\left(x\boldsymbol{Q}_c\right)$.

We know from [7, p. 46] that the expected time until absorption can be written as

$$\mathrm{E}\left[T_c(i,j)\right] = -\boldsymbol{e}_{|\mathcal{T}_c|}^{\mathrm{ind}(i,j)} \cdot \left(\boldsymbol{Q}_c\right)^{-1} \cdot \mathbf{1}_{|\mathcal{T}_c|}, \quad (i,j) \in \mathcal{T}_c, \tag{2}$$

where the existence of $\left(\boldsymbol{Q}_c\right)^{-1}$ is a consequence of the fact that all states in $\mathcal{T}_c$ are transient [7, p. 45]. Inverting $\boldsymbol{Q}_c$ analytically can rapidly become cumbersome as s or r increases. We will instead perform numerical computations as reported in Section 6. Consider now

$$T_c((i,j),(i',j')) := \int_0^{T_c(i,j)} \mathbb{1}\{(X_c(t), Y_c(t)) = (i',j')\}\, dt$$

that is the total time spent by the CTMC in transient state (i',j') given that $\{X_c(0), Y_c(0)\} = (i,j)$. It can also be shown that [4, p. 419]

$$\mathrm{E}\left[T_c((i,j),(i',j'))\right] = -\boldsymbol{e}_{|\mathcal{T}_c|}^{\mathrm{ind}(i,j)} \cdot \left(\boldsymbol{Q}_c\right)^{-1} \cdot {}^t\boldsymbol{e}_{|\mathcal{T}_c|}^{\mathrm{ind}(i',j')}, \quad (i,j),(i',j') \in \mathcal{T}_c, \tag{3}$$

where ${}^t\boldsymbol{y}$ denotes the transpose of a given vector $\boldsymbol{y}$. In other words, $\mathrm{E}\left[T_c((i,j),(i',j'))\right]$ is the entry of matrix $(-\boldsymbol{Q}_c)^{-1}$ at row $\mathrm{ind}(i,j)$ and column $\mathrm{ind}(i',j')$.

4.2 Data Availability

In this section we introduce different metrics to quantify the availability of D. We are interested in the fraction of time spent by the CTMC in any given state (i',j') before absorption. However, this quantity is difficult to find in closed-form. Therefore, we resort to using the following approximation

$$\mathrm{E}\left[\frac{T_c((i,j),(i',j'))}{T_c(i,j)}\right] \approx \frac{\mathrm{E}[T_c((i,j),(i',j'))]}{\mathrm{E}[T_c(i,j)]}. \tag{4}$$

Here, (i,j) is the state of D at $t=0$. This approximation have been validated through simulations, as shown later in Section 6. With this approximation in mind, we introduce two availability metrics: the first can be interpreted as the expected number of fragments of D that are in the system during the lifetime of D; the second can be interpreted as the fraction of time when at least m fragments are in the system during the lifetime of D. More formally, given that $(X_c(0),Y_c(0)) = (i,j) \in \mathcal{T}_c$, we define

$$M_{c,1}(i,j) := \sum_{(i',j')\in\mathcal{T}_c} i' \frac{\mathrm{E}[T_c((i,j),(i',j'))]}{\mathrm{E}[T_c(i,j)]}, \tag{5}$$

$$M_{c,2}((i,j),m) := \sum_{(i',j')\in\mathcal{T}_c, i'\geq m} \frac{\mathrm{E}[T_c((i,j),(i',j'))]}{\mathrm{E}[T_c(i,j)]}. \tag{6}$$

5 Distributed Repair Systems

In this section, we model P2P storage systems that implement a distributed recovery mechanism. According to the description and assumptions listed in Section 3, the state of data D can be modeled by an absorbing Markov chain $\{(X_d(t),Y_d(t)) : t \geq 0\}$, where $X_d(t)$ and $Y_d(t)$ denote respectively the number of fragments in the system that are available for download and the state of the recovery process. Unlike the centralized scheme, the distributed scheme repairs fragments only one at a time. Therefore, $X_d(t)$ takes value in the set $\{s-1, s, \ldots, s+r\}$. There are only s stages in the recovery process that correspond each to an exponential with it own rate (i.e., $Y_d(t) \in \{0,1,\ldots,s-1\}$). As in Section 4, D is available when $X_d(t) \geq s$, unavailable when $X_d(t) < s$ but $X_d(t)+Y_d(t) \geq s$, and lost otherwise (situation modeled by a single absorbing state a). The set of transient states $\mathcal{T}_d$ is

$$\begin{array}{lll}
\mathcal{T}_d := \{ & (s-1,1),(s-1,2),\ldots,(s-1,s-1), & \} \quad D \text{ is unavailable} \\
 & \left.\begin{array}{l}
(s,0),(s,1),\ldots,(s,s-1), \\
(s+1,0),(s+1,1),\ldots,(s+1,s-1),\ldots, \\
(s+r-1,0),(s+r-1,1),\ldots,(s+r-1,s-1), \\
(s+r,0)\ \}
\end{array}\right\} & D \text{ is available} \\
|\mathcal{T}_d| = s(r+1). &
\end{array}$$

The analysis of the absorbing Markov chain $\{(X_d(t),Y_d(t)) : t \geq 0\}$ that takes value in $\mathcal{T}_d \cup \{a\}$ is very similar to the analysis in Section 4, we will then only sketch it. In particular, $\boldsymbol{R}_d$ and $\boldsymbol{Q}_d$ have similar definitions as $\boldsymbol{R}_c$ and $\boldsymbol{Q}_c$ after replacing the

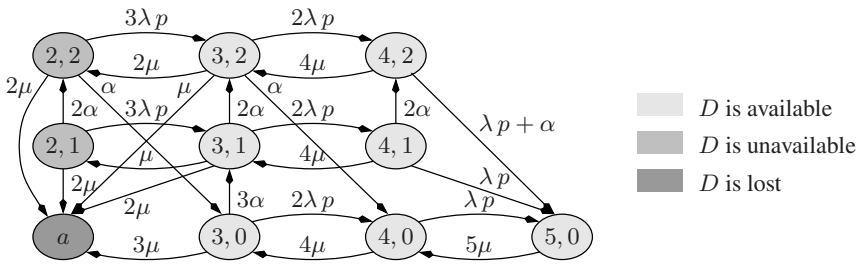

Figure 2: The Markov chain $\{(X_d(t), Y_d(t)),\ t \geq 0\}$ when $s = 3$, $r = 2$, and $k = 2$.

subscript "c" with the subscript "d" whenever needed. The non-zero elements of $\boldsymbol{R}_d$ and $\boldsymbol{Q}_d$ are as follows

$$r_d(s-1, j) = (s-1)\mu, \text{ for } j = 1, \ldots, s-1; \quad r_d(s, j) = (s-j)\mu, \text{ for } j = 0, \ldots, s-1.$$

$$q_d((i,j),(i-1,j)) = \begin{cases} j\mu, & \text{for} \quad i = s, j = 1, \ldots, s-1; \\ i\mu, & \text{for} \quad i = s+1, \ldots, s+r-1, j = 0, \ldots, s-1; \\ & \text{or} \quad i = s+r, j = 0. \end{cases}$$

$$q_d((i,j),(i,j+1)) = (s-j)\alpha, \quad \text{for} \quad i = s, \ldots, s+r-k, j = 0; \quad \text{or} \quad i = s-1, \ldots, s+r-1, j = 1, \ldots, s-2.$$

$$q_d((i,s-1),(i+1,0)) = \alpha, \quad \text{for} \quad i = s-1, \ldots, s+r-2.$$

$$q_d((i,j),(i+1,j)) = (s+r-i)\lambda p, \quad \text{for} \quad i = s-1, j = 1, \ldots, s-1; \quad \text{or} \quad i = s, \ldots, s+r-2, j = 0, \ldots, s-1.$$

$$q_d((s+r-1,j),(s+r,0)) = \lambda p + \mathbb{1}\{j = s-1\}\,\alpha, \quad \text{for} \quad j = 0, \ldots, s-1.$$

$$q_d((i,j),(i,j)) = -r_d(i,j) - \textstyle\sum_{(i',j') \in \mathcal{T}_d - \{(i,j)\}} q_d((i,j),(i',j')), \quad \text{for } (i,j) \in \mathcal{T}_d.$$

For illustration purposes, we depict in Fig. 2 an example of the absorbing CTMC with its non-zero transition rates when $s = 3$, $r = 2$, and $k = 2$.

We can now derive closed-form expressions for the distribution of the conditional block lifetime, its expectation, and the two availability metrics, as was done in Section 4, by simply replacing in (1), (2), (3), (5) and (6) the subscript "c" with the subscript "d". Alike for the centralized case, we will perform numerical computations as it is not tractable to explicitly invert $\boldsymbol{Q}_d$.

6 Numerical Results

In this section, we first validate the approximation made in (4) which has been made to compute the two availability metrics, then proceed with the presentations of the numerical computations. Throughout this section, we consider two sets of parameters that correspond each to a particular context. In the "PlanetLab" context, we set $1/\lambda = 61$ hours and $1/\mu = 181$ hours according to [9], $p = 0.3$ to reflect that disconnections are most likely due to software or hardware problems, and we vary the redundancy r from 1 until s. In the "Internet" context, peers churn rate is much higher [2] (namely, hosts join and leave the system 6.4 times a day on average), which led us to set $1/\lambda = 1$ hours, $1/\mu = 3$ hours, $p = 0.7$, and to vary r from 1 to $2s$. In both contexts, we let $s = 8$, $k = 1, \ldots, r$. We assume the peers' upload capacity that is dedicated to a single connection to be 10kbps (cf. [5]), and the total upload capacity of the central authority

to be 400kbps. Hence, considering fragments of size 1MB, we obtain $1/\alpha = 838.8608$ seconds, $1/\beta = 20.97152 \times s$ seconds.

6.1 Validation of (4)

To this end, we have simulated the CTMC $\{X_d(t), Y_d(t) : t \geq 0\}$ in the "Internet" context but have varied r from 1 to 4, yielding a total of 10 different simulation scenarios. We estimate the left-hand side of (4) by averaging the corresponding simulated results. In order to obtain a maximum estimation error of about 1% with 97% confidence interval, we need to average over 150 sampled values. Hence, each simulation is repeated 150 times. In each simulation, we have a total of $|\mathcal{T}_d| = 8(r+1)$ instances of (4), according to the possible choices of the initial state, yielding a total of $\sum_{r=1}^{4} 8(r+1)r = 320$ different instances. For each instance, we compute the relative error between the estimation of the left-hand side of (4) (the "correct" value) and the right-hand side of (4) (the approximate value, computed analytically using (2)-(3)).

The empirical complementary cumulative distribution function of the relative error is displayed in Fig. 3. As can be seen, we have found only 10% of the values that are larger than 0.9×10^{-3} and, most importantly, the maximum value of the relative error is 0.0028. *We conclude that the approximation* (4) *is very good and will definitely not imperil the correctness of any computation based on it.*

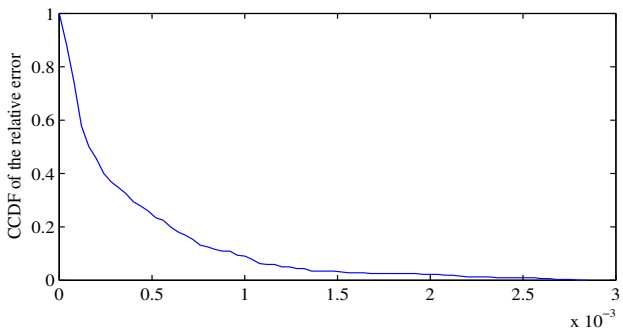

Figure 3: The CCDF of the relative error induced by the approximation (4)

6.2 Performance analysis

We have solved numerically (1), (2), (3), (5) and (6) given that all $s + r$ fragments of D are initially available, considering either Internet or PlanetLab context, and either the centralized or distributed recovery scheme. Results are reported partially in Table 1. It appears that, whichever the scenario or the recovery mechanism considered, the expected data lifetime increases roughly exponentially with r and decreases with an increasing k. Regardless of the context considered, the distributed scheme yields a significantly smaller expected data lifetime than the centralized scheme, especially when the storage overhead, r/s, is high; cf. columns 3-4 in Table 1. The difference in performance is more pronounced in the Internet context. Regarding the expected number of available fragments, we again observe that the distributed scheme is less efficient than the centralized one. Observe how the performance deteriorates as peer churn becomes

Table 1: Expected lifetime and first availability metric

Internet context		$\mathrm{E}[T(s+r,0)]$ (in days)		$M_1(s+r,0)$	
	$s=8$	cent. repair	dist. repair	cent. repair	dist. repair
$k=1$	$r/s=1/2$	0.14	0.08	9.91	8.99
	$r/s=1$	0.85	0.35	11.72	10.58
	$r/s=3/2$	14.49	2.25	14.46	12.40
	$r/s=2$	105.80	23.61	16.55	14.73
$k=2$	$r/s=1/2$	0.13	0.08	9.90	8.69
	$r/s=1$	0.81	0.35	11.70	10.27
	$r/s=3/2$	13.90	2.25	14.46	12.10
	$r/s=2$	101.70	23.61	16.53	14.43
$k=4$	$r/s=1$	0.71	0.33	11.66	9.67
	$r/s=3/2$	12.38	2.23	14.45	11.50
	$r/s=2$	91.59	23.58	16.48	13.83
$k=8$	$r/s=2$	64.16	22.31	16.30	12.61
PlanetLab context		$\mathrm{E}[T(s+r,0)]$ (in months)		$M_1(s+r,0)$	
	$s=8$	cent. repair	dist. repair	cent. repair	dist. repair
$k=1$	$r/s=1/4$	0.32	0.11	7.81	8.04
	$r/s=1/2$	2.15	1.05	11.01	8.68
	$r/s=3/4$	17.12	7.61	13.18	9.80
	$r/s=1$	262.16	46.24	15.11	12.12
$k=2$	$r/s=1/2$	0.81	0.37	10.34	8.19
	$r/s=3/4$	6.95	3.20	12.76	9.25
	$r/s=1$	110.03	23.34	14.72	11.37
$k=4$	$r/s=1$	13.33	4.34	13.77	9.81

more important: compare for instance in Table 1 rows 4 vs. 20, and 11 vs. 24 (these correspond to the same storage overhead and the same value of k). This is particularly true for the distributed recovery mechanism. We conclude that *when peers churn rate is high, only the centralized repair scheme can be efficient should the storage overhead be kept within a reasonable value (that is $r/s \leq 2$).* As the distributed repair scheme is more scalable than the centralized one, it will be a good implementation choice in large networks where hosts have a good availability.

6.3 Setting the system's key parameters

We illustrate now how our models can be used to set the system parameters r and k such that predefined requirements on data lifetime and availability are fulfilled. We assume the recovery mechanism is centralized and the context is similar to PlanetLab. We have picked one to two contour lines of each of the performance metrics studied in this report and report them in Fig. 6.3. Consider point A which corresponds to $r = 6$ and $k = 1$ (recall $s = 8$). Selecting this point as the operating point of the P2PSS ensures (roughly) the following: given that each data is initiated with $s + r$ available fragments, then (i) the expected data lifetime is 18 months; (ii) only 11% of the stored data would be lost after 3 months; (iii) as long as D is not lost, 13 fragments of D are expected to be in the system; (iv) during 99.7% of its lifetime, D is available for

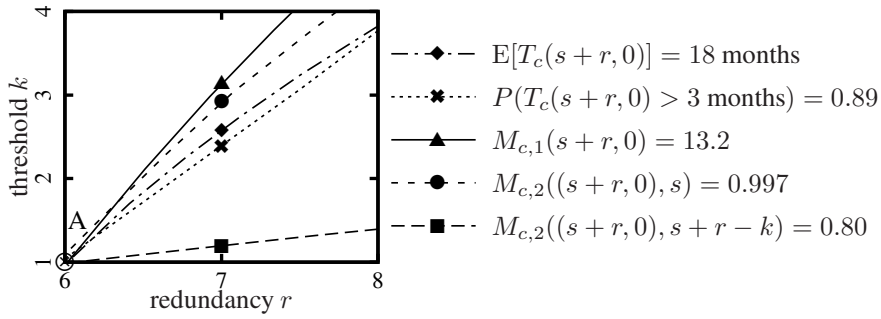

Figure 4: Contour lines of performance metrics (PlanetLab context, centralized repair).

download; and (v) during 80% of the lifetime of D, at least $s + r - k = 13$ fragments of D are available for download in the system. Observe that the storage overhead, r/s, is equal to 0.75.

6.4 Impact of the size of fragments

Given the size of data D, a larger size of fragments translates into a smaller s. We have computed all pairs (r, k) with $s = 8$ and $s = 16$ that ensure $P(T_c(s + r, 0) > 3 \text{ months}) = 0.89$ in the PlanetLab context, i.e., only 11% of the total data would be lost after 3 months. In particular, operating points $r = 6$ and $k = 1$ with $s = 8$, and $r = 12$ and $k = 7$ with $s = 16$ satisfy the above requirement, and additionally yield the same storage overhead (namely, 0.75). But, and this is important, the former point invokes the recovery process much more often (and potentially unnecessarily) than the latter point, suggesting that *large fragments size reduces the efficiency of the recovery mechanism*. This observation should be moderated by the fact that fragments size when $s = 8$ is twice their size when $s = 16$, yielding a different bandwidth usage per recovery. We currently cannot say how does the bandwidth usage per recovery vary with the size of fragments. However, we know for sure that its effect will not be the same in both centralized and distributed schemes because of the additional upload stages in the centralized implementation. A careful analysis of this issue is the objective of ongoing research.

7 Conclusion

We have proposed analytical models for evaluating the performance of two approaches for recovering lost data in distributed storage systems. We have analyzed the lifetime and the availability of data achieved by both centralized- and distributed-repair systems through Markovian analysis considering realistic assumptions. Numerical computations have been undertaken to illustrate several issues on the performance. We conclude that, using our theoretical framework, it is easy to tune and optimize the system parameters for fulfilling predefined requirements. We find that, in stable environments such as local area or research laboratory networks where machines are usually highly available, the distributed-repair scheme offers a reliable, scalable and cheap storage/backup solution. This is in contrast with the case of highly dynamic environments, where the

distributed-repair scheme is inefficient as long as the storage overhead is kept reasonable. P2PSS with centralized-repair scheme are efficient in any environment but have the disadvantage of relying on a centralized authority. Our analysis also suggests that the use of large size fragments reduces the efficiency of the recovery mechanism.

References

[1] S. Alouf, A. Dandoush, and P. Nain. Performance analysis of peer-to-peer storage systems. In *Proc. of 20th ITC*, volume 4516 of *Lecture Notes in Computer Science*, pages 642–653, Ottawa, Canada, June 2007.

[2] R. Bhagwan, S. Savage, and G. Voelker. Understanding availability. In *Proc. of 2nd IPTPS*, Berkeley, California, February 2003.

[3] R. Bhagwan, K. Tati, Y. Cheng, S. Savage, and G.M. Voelker. Total Recall: System support for automated availability management. In *Proc. of ACM/USENIX NSDI '04*, pages 337–350, San Francisco, California, March 2004.

[4] C. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, 1997.

[5] A. Guha, N. Daswani, and R. Jain. An experimental study of the skype peer-to-peer VoIP system. In *Proc. of 5th IPTPS*, Santa Barbara, California, February 2006.

[6] P. Harrison and S. Zertal. Queueing models of RAID systems with maxima of waiting times. *Performance Evaluation Journal*, 64(7-8):664–689, August 2007.

[7] M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models. An Algorithmic Approach*. John Hopkins University Press, Baltimore, 1981.

[8] PlanetLab. An open platform for developing, deploying, and accessing planetary-scale services. `http://www.planet-lab.org/`, 2007.

[9] S. Ramabhadran and J. Pasquale. Analysis of long-running replicated systems. In *Proc. of IEEE Infocom '06*, Barcelona, Spain, April 2006.

[10] I.S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of SIAM*, 8(2):300–304, June 1960.

[11] A. Rowstron and P. Druschel. Storage management and caching in PAST, a large-scale, persistent peer-to-peer storage utility. In *Proc. of ACM SOSP '01*, pages 188–201, Banff, Canada, October 2001.

[12] J. Stribling. PlanetLab - All Pairs Pings. `http://pdos.csail.mit.edu/~strib/pl_app`, 2005.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399