Sauvegarde collaborative en pair-à-pair

Fabrice Le Fessant
Fabrice.Le_Fessant@inria.fr

ASAP Team
INRIA Saclay – Île de France

Octobre 2008

Plan

INRIA

Plan

INRIA

Définition

La sauvegarde collaborative ou backup en pair-à-pair

Utiliser l'espace disque libre sur d'autres ordinateurs connectés au réseau pour sauvegarder ses propres données.

Trois contextes aux difficultés croissantes :

- Sauvegarder l'ordinateur de mes parents sur mes ordinateurs
- Sauvegarder les données de l'entreprise sur son réseau interne
- Sauvegarder mes données sur Internet

INRIA

Définition

La sauvegarde collaborative ou backup en pair-à-pair

Utiliser l'espace disque libre sur d'autres ordinateurs connectés au réseau pour sauvegarder ses propres données.

Trois contextes aux difficultés croissantes :

- Sauvegarder l'ordinateur de mes parents sur mes ordinateurs
- Sauvegarder les données de l'entreprise sur son réseau interne
- Sauvegarder mes données sur Internet

INRIA

L'utilité de la sauvegarde collaborative

De plus en plus de données numériques non protégées

- Particuliers : couriels, contenus générés (photos, films, blogs)
- Professionnels : utilisation des portables pour la mobilité

De plus en plus d'espace disque connecté :

- Prolifération des connexions haut-débit : ADSL, fibre, 3G, ...
- Augmentation de la capacité des disques : 3 Go en 1995, 160 Go en 2000, 1 To en 2008. Idem pour les portables.
- Et l'espace libre : 1% de 1To = 10Go = 5000 photos.

Les technologies sont au rendez-vous :

- Cryptographie : chiffrement des données, codes correcteurs, challenges de possession
- Pair-à-pair : organisation automatique du réseau, bande passante

L'utilité de la sauvegarde collaborative

De plus en plus de données numériques non protégées

- Particuliers : couriels, contenus générés (photos, films, blogs)
- Professionnels : utilisation des portables pour la mobilité

De plus en plus d'espace disque connecté :

- Prolifération des connexions haut-débit : ADSL, fibre, 3G, ...
- Augmentation de la capacité des disques : 3 Go en 1995, 160 Go en 2000, 1 To en 2008. Idem pour les portables.
- Et l'espace libre : 1% de 1To = 10Go = 5000 photos.

Les technologies sont au rendez-vous :

- Cryptographie : chiffrement des données, codes correcteurs, challenges de possession
- Pair-à-pair : organisation automatique du réseau, bande passante

L'utilité de la sauvegarde collaborative

De plus en plus de données numériques non protégées

- Particuliers : couriels, contenus générés (photos, films, blogs)
- Professionnels : utilisation des portables pour la mobilité

De plus en plus d'espace disque connecté :

- Prolifération des connexions haut-débit : ADSL, fibre, 3G, ...
- Augmentation de la capacité des disques : 3 Go en 1995, 160 Go en 2000, 1 To en 2008. Idem pour les portables.
- Et l'espace libre : 1% de 1To = 10Go = 5000 photos.

Les technologies sont au rendez-vous :

- Cryptographie : chiffrement des données, codes correcteurs, challenges de possession
- Pair-à-pair : organisation automatique du réseau, bande passante

Les atouts de la sauvegarde collaborative

De nombreux avantages

- Facilité : pas de matériel particulier, configuration rudimentaire
- Rapididté : sauvegarde et restauration dès qu'on est connecté
- Confidentialité : chiffrement des données
- Résistance : le système se surveille et se corrige automatiquement
- Distance : une catastrophe locale ne met pas en danger les données

Un inconvénient :

Pas de garantie de récupération des données
$\Rightarrow$ en complément d'autres techniques de sauvegarde...

Les alternatives

Pourquoi une sauvegarde ?

- 60 % des utilisateurs n'ont pas de sauvegarde
- 60 % des compagnies qui perdent leur données font faillite dans les 6 mois

Les autres sauvegardes

- Sur support passif (bandes, disques) : compliqué, vieillissement
- Sur support actif (disque externe) : compliqué, panne, vol, incendie
- Sur support distant (serveurs distants) : faillites, sécurité, bugs

INRIA

Les alternatives

Pourquoi une sauvegarde ?

- 60 % des utilisateurs n'ont pas de sauvegarde
- 60 % des compagnies qui perdent leur données font faillite dans les 6 mois

Les autres sauvegardes

- Sur support passif (bandes, disques) : compliqué, vieillissement
- Sur support actif (disque externe) : compliqué, panne, vol, incendie
- Sur support distant (serveurs distants) : faillites, sécurité, bugs

INRIA

Peer-to-Peer versus Cloud Computing

Deux approches opposées

- Cloud Computing : un service (payant) qui croît et décroît en fonction des besoins de ses utilisateurs (data-center)
- Peer-to-Peer : un service (gratuit) constitué des ressources fournies par ses utilisateurs

Le Cloud Computing va-t-il tout résoudre ?

- Amazon, Google, Flickr, Facebook ont des centaines de millions d'utilisateurs
- Mais :
  - Logiciel Propriétaire -> Logiciel Libre -> Stockage Propriétaire
  - Pas d'intéropérabilité (kidnapping des données), dispersion des données, pas de confidentialité (vie privée), boîtes noires (sécurité, autres services)

Peer-to-Peer versus Cloud Computing

Deux approches opposées

- Cloud Computing : un service (payant) qui croît et décroît en fonction des besoins de ses utilisateurs (data-center)
- Peer-to-Peer : un service (gratuit) constitué des ressources fournies par ses utilisateurs

Le Cloud Computing va-t-il tout résoudre ?

- Amazon, Google, Flickr, Facebook ont des centaines de millions d'utilisateurs
- Mais :
  - Logiciel Propriétaire -> Logiciel Libre -> Stockage Propriétaire
  - Pas d'intéropérabilité (kidnapping des données), dispersion des données, pas de confidentialité (vie privée), boîtes noires (sécurité, autres services)

Plan

INRIA

Fonctionnement

La sauvegarde

- Détection des modifications du système de fichiers
- Extraction des données à sauvegarder
- Chiffrement des données
- Constitution d'archives de fichiers
- Sélection des points de stockage
- Transferts des archives et index à distance

La restauration

- Récupération des points de stockage
- Récupération des listes de fichiers
- Téléchargement des archives
- Extraction des fichiers

Fonctionnement

La sauvegarde

- Détection des modifications du système de fichiers
- Extraction des données à sauvegarder
- Chiffrement des données
- Constitution d'archives de fichiers
- Sélection des points de stockage
- Transferts des archives et index à distance

La restauration

- Récupération des points de stockage
- Récupération des listes de fichiers
- Téléchargement des archives
- Extraction des fichiers

Confidentialité des données

La nécessité du chiffrement

- Les données sauvegardées peuvent être :
  - Interceptées sur le réseau
  - examinées par le propriétaire d'un point de stockage
- Aucune donnée n'est transmise sur le réseau sans chiffrement préalable

Méthode de chiffrement

- Chaque utilisateur possède une paire de clés asymétriques
- Création d'une clé de session par archive
- Chiffrement asymétrique de la clé de session (RSA)
- Auto-Chiffrement des fichiers (AES sur hash du fichier)
- Chiffrement symmétrique des hashs de fichiers

Confidentialité des données

La nécessité du chiffrement

- Les données sauvegardées peuvent être :
  - Interceptées sur le réseau
  - examinées par le propriétaire d'un point de stockage
- Aucune donnée n'est transmise sur le réseau sans chiffrement préalable

Méthode de chiffrement

- Chaque utilisateur possède une paire de clés asymétriques
- Création d'une clé de session par archive
- Chiffrement asymétrique de la clé de session (RSA)
- Auto-Chiffrement des fichiers (AES sur hash du fichier)
- Chiffrement symmétrique des hashs de fichiers

Réplication versus Codes Correcteurs

Tolérer les pannes des points de stockage

Exemple : coût pour protéger une archive de 100 Mo contre 9 pannes

Réplication : simples copies

- Le système crée 10 copies des 100 Mo sur 10 pairs
- Coût total : 1 Go, coût réparation : 100 Mo

Codes correcteurs : combinaisons linéaires

- Le système découpe l'archive en 10 blocs de 10 Mo
- Ajout de 9 blocs de 10 Mo, combinaisons linéaires des précédents
- Les 19 blocs sont répartis sur 19 pairs
- Coût total : 190 Mo, coût réparation : 100 Mo

Réplication versus Codes Correcteurs

Tolérer les pannes des points de stockage

Exemple : coût pour protéger une archive de 100 Mo contre 9 pannes

Réplication : simples copies

- Le système crée 10 copies des 100 Mo sur 10 pairs
- Coût total : 1 Go, coût réparation : 100 Mo

Codes correcteurs : combinaisons linéaires

- Le système découpe l'archive en 10 blocs de 10 Mo
- Ajout de 9 blocs de 10 Mo, combinaisons linéaires des précédents
- Les 19 blocs sont répartis sur 19 pairs
- Coût total : 190 Mo, coût réparation : 100 Mo

Réplication versus Codes Correcteurs

Tolérer les pannes des points de stockage

Exemple : coût pour protéger une archive de 100 Mo contre 9 pannes

Réplication : simples copies

- Le système crée 10 copies des 100 Mo sur 10 pairs
- Coût total : 1 Go, coût réparation : 100 Mo

Codes correcteurs : combinaisons linéaires

- Le système découpe l'archive en 10 blocs de 10 Mo
- Ajout de 9 blocs de 10 Mo, combinaisons linéaires des précédents
- Les 19 blocs sont répartis sur 19 pairs
- Coût total : 190 Mo, coût réparation : 100 Mo

Dégradation de la sauvegarde

Le nombre de réplicas décroît en permanence :

- Déconnexions définitives des pairs à remplacer
- Déconnexions temporaires des pairs à ne pas remplacer !
- Pairs malicieux ne stockant pas les données

Danger de pertes de données :

- Trop peu de réplicas : impossible de restaurer les données
  $\Rightarrow$ réparation régulière des réplicas

INRIA

Dégradation de la sauvegarde

Le nombre de réplicas décroît en permanence :

- Déconnexions définitives des pairs à remplacer
- Déconnexions temporaires des pairs à ne pas remplacer !
- Pairs malicieux ne stockant pas les données

Danger de pertes de données :

- Trop peu de réplicas : impossible de restaurer les données
  $\Rightarrow$ réparation régulière des réplicas

INRIA

Maintenance de la sauvegarde

Surveillance des points de stockage

- Le système doit observer la disponibilité des pairs stockant les données en permanence
- En cas de risque, le système doit rétablir la redondance
  $\Rightarrow$ Utilisation de codes correcteurs minimisant le coût des réparations

Vérification des données

- Envoyer régulièrement des challenges pour vérifier que les pairs n'ont pas effacé les données stockées
  $\Rightarrow$ Infinité de challenges cryptographiques générées à partir d'une signature de chaque bloc et de fonctions homomorphiques.

INRIA

Maintenance de la sauvegarde

Surveillance des points de stockage

- Le système doit observer la disponibilité des pairs stockant les données en permanence
- En cas de risque, le système doit rétablir la redondance
  $\Rightarrow$ Utilisation de codes correcteurs minimisant le coût des réparations

Vérification des données

- Envoyer régulièrement des challenges pour vérifier que les pairs n'ont pas effacé les données stockées
  $\Rightarrow$ Infinité de challenges cryptographiques générées à partir d'une signature de chaque bloc et de fonctions homomorphiques.

INRIA

Plan

INRIA

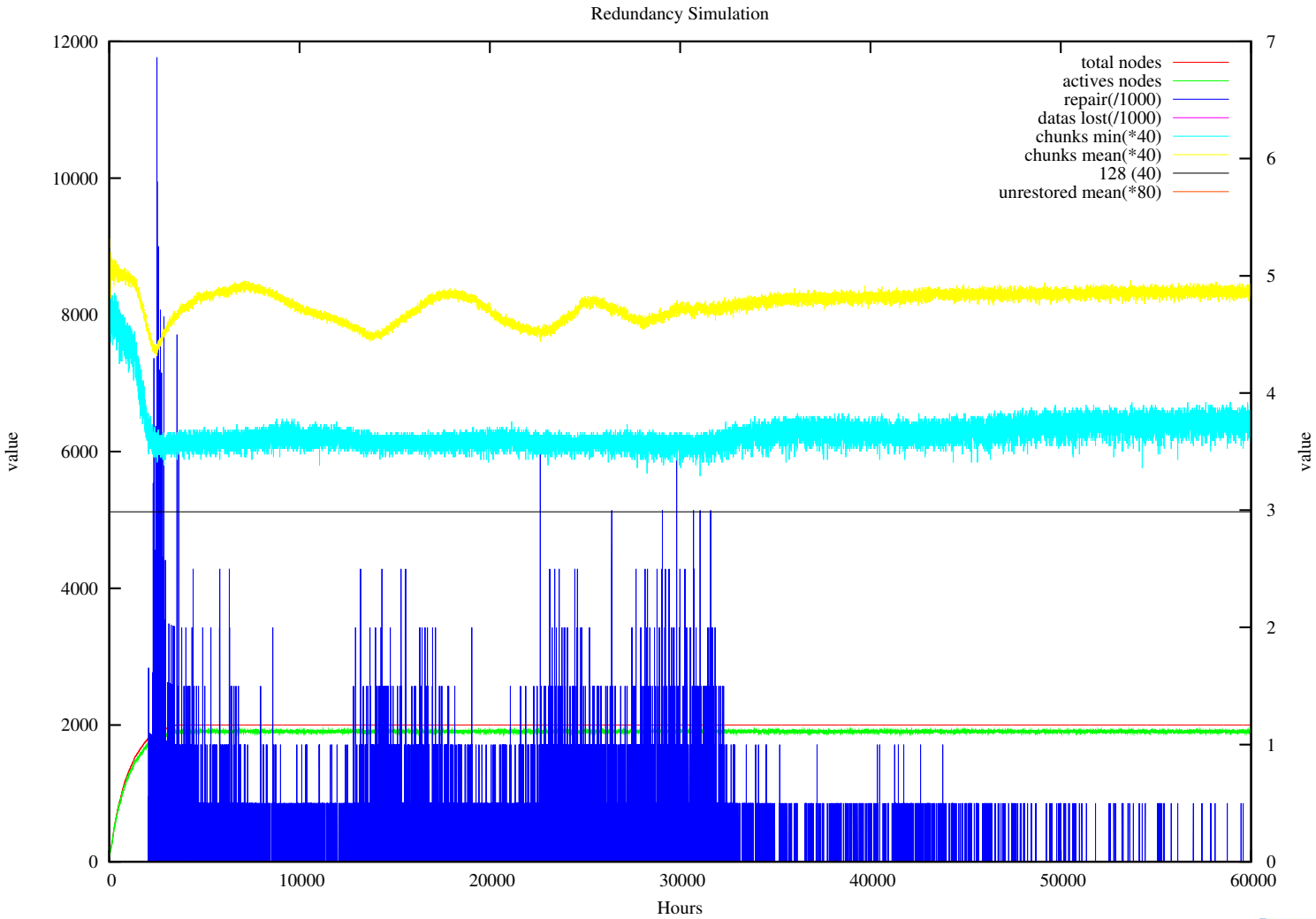
Simulations

Simulation d'une sauvegarde collaborative

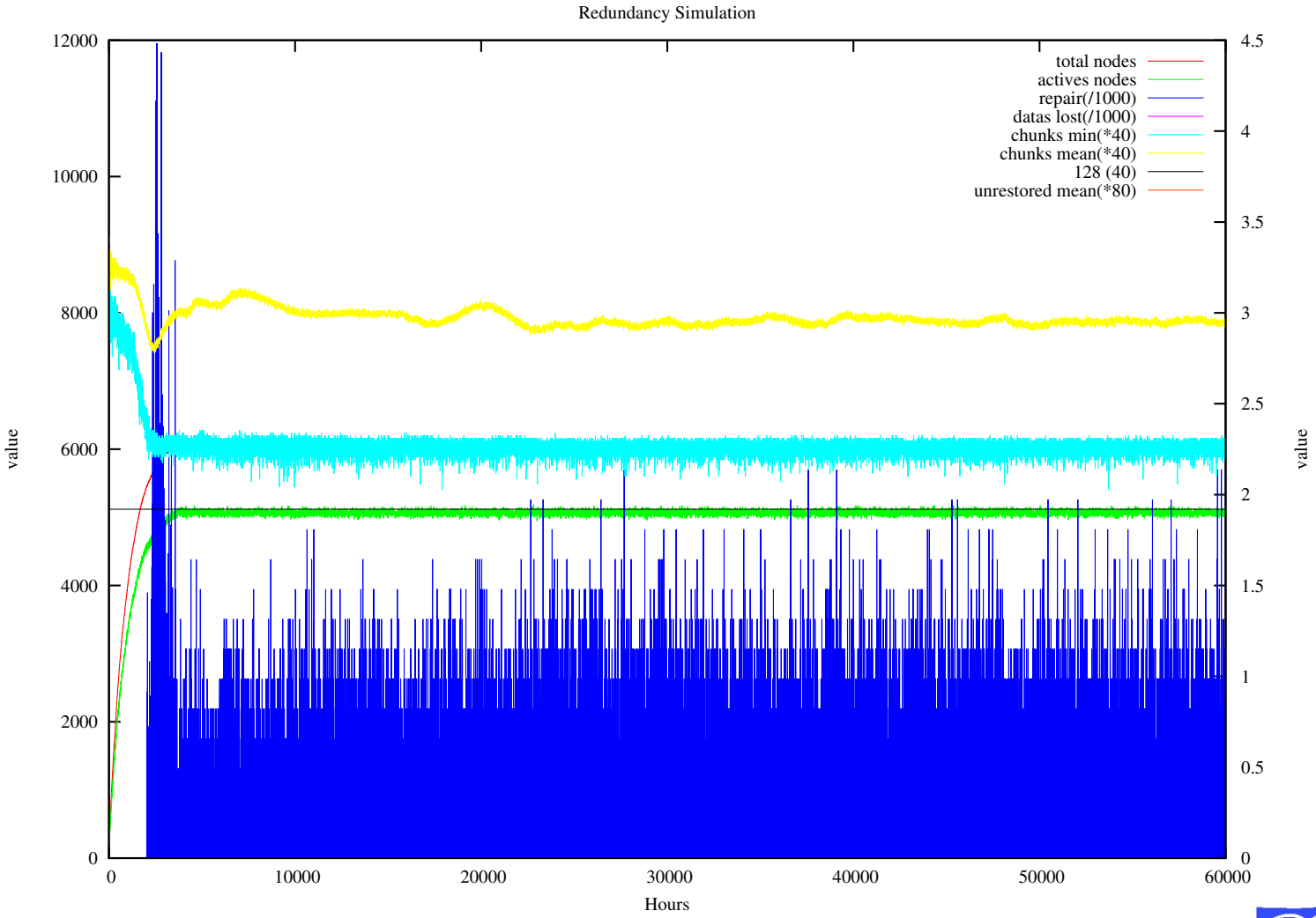
- Pairs triés suivant leur participation au système
  $\Rightarrow$ Mesure fiable de la disponibilité
- Incitations à participer :
  - Les pairs les plus stables préfèrent les pairs les plus stables
  - Les nouveaux pairs doivent travailler plus, mais pas trop
- Mesure du nombre de réparations pour 1000 heures (41 jours)

INRIA

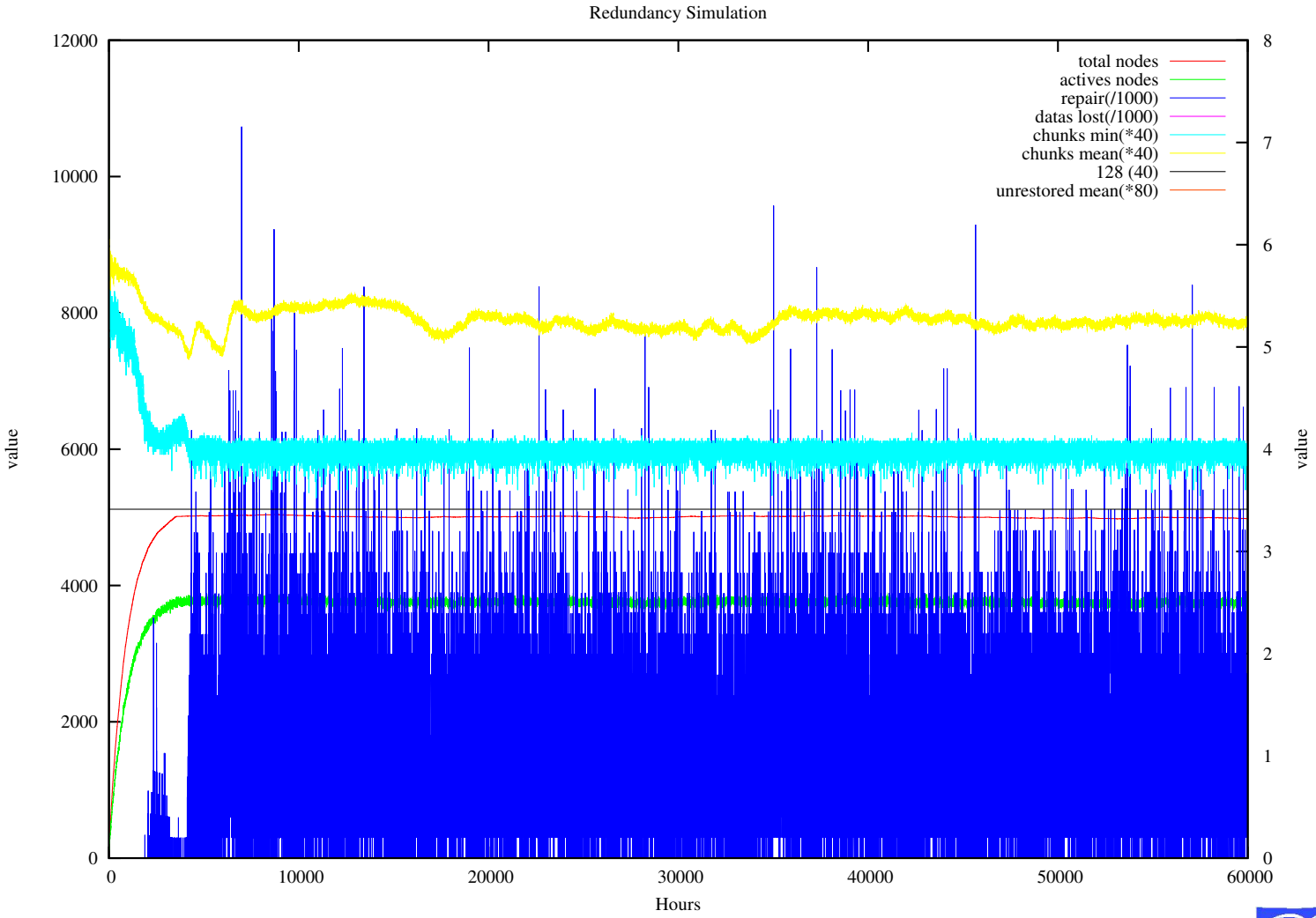
Pairs les plus stables



Pairs moins stables



Pairs les moins stables



Plan

INRIA

Conclusion

Une application prometteuse

- Un besoin important pour les particuliers/professionnels
- Une technologie à contre-courant du Cloud Computing
- Pas ou peu deployée : beaucoup de problèmes pour passer de la théorie à la pratique
- Réhabilitation du pair-à-pair

INRIA