Preservação da Web através de replicação distribuída em larga escala

André Nogueira

AWP – Arquivo da Web Portuguesa - http://arquivo-web.fccn.pt/

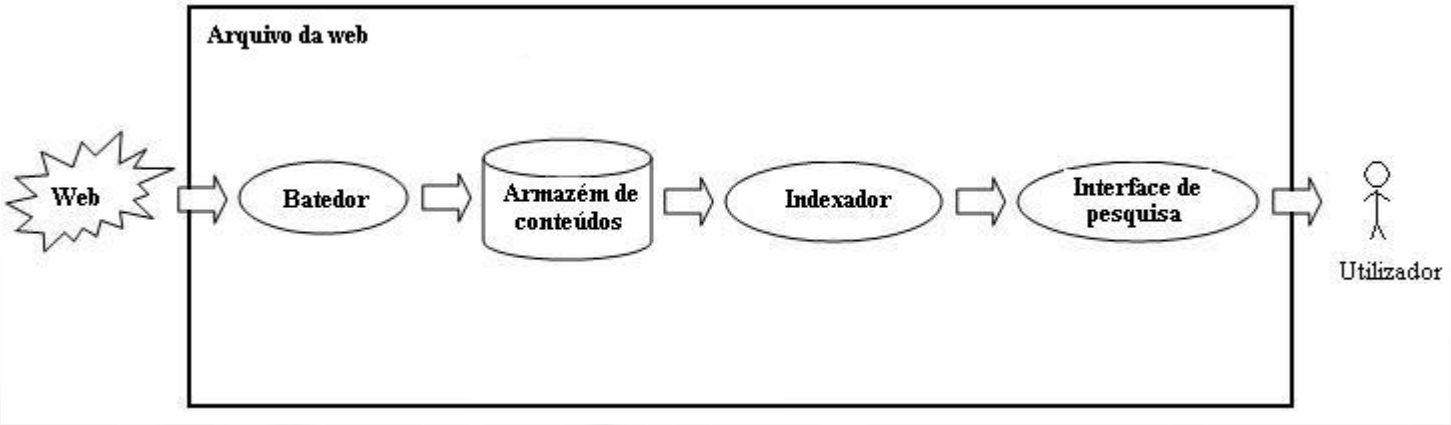
andre [ponto] nogueira [arroba] fccn [ponto] pt

ARQUIVO DA WEB PORTUGUESA

Introdução

- A Web é a maior fonte de informação construída
  - Jornais, livros, documentação técnica, blogs
  - Informação publicada exclusivaente na Web
- A informação na Web é efémera
- Preservação iniciada pelos arquivos da Web
- Primeira iniciativa: Internet Archive

Arquitectura geral



Problemas de armazenamento

- Armazenamento centralizado
  - Incêndio ou inundação das instalações
- Cópias de segurança → Duplicação de custos
- Recolha contínua → Aumento do espaço de armazenamento
- Resultados de uma recolha do AWP
  - Sítios Web visitados: 455 mil
  - Tamanho máximo ficheiro: 10 MBytes
  - Total recolhido: 2,5 TBytes

Solução possível?

- Em 2015 haverá mais de 1000 milhões de computadores ligados à rede mundial

  BOINC: A System for Public-Resource Computing and Storage. David P. Anderson. 5th IEEE/ACM International Workshop on Grid Computing. November 8, 2004, Pittsburgh, USA.

- Actualmente, um PC tem 100 GBytes de capacidade de disco
- Se 100 milhões de utilizadores disponibilizarem 10% (10 GBytes) de espaço de disco → Total disponível seria 1 ExaByte (10^{18} bytes)

Replicador de ficheiros Arc (rArc)

Tópicos

- Requisitos
- Trabalho Relacionado
- Arquitectura
- Funcionamento
- Avaliação
- Trabalho Futuro
- Conclusões

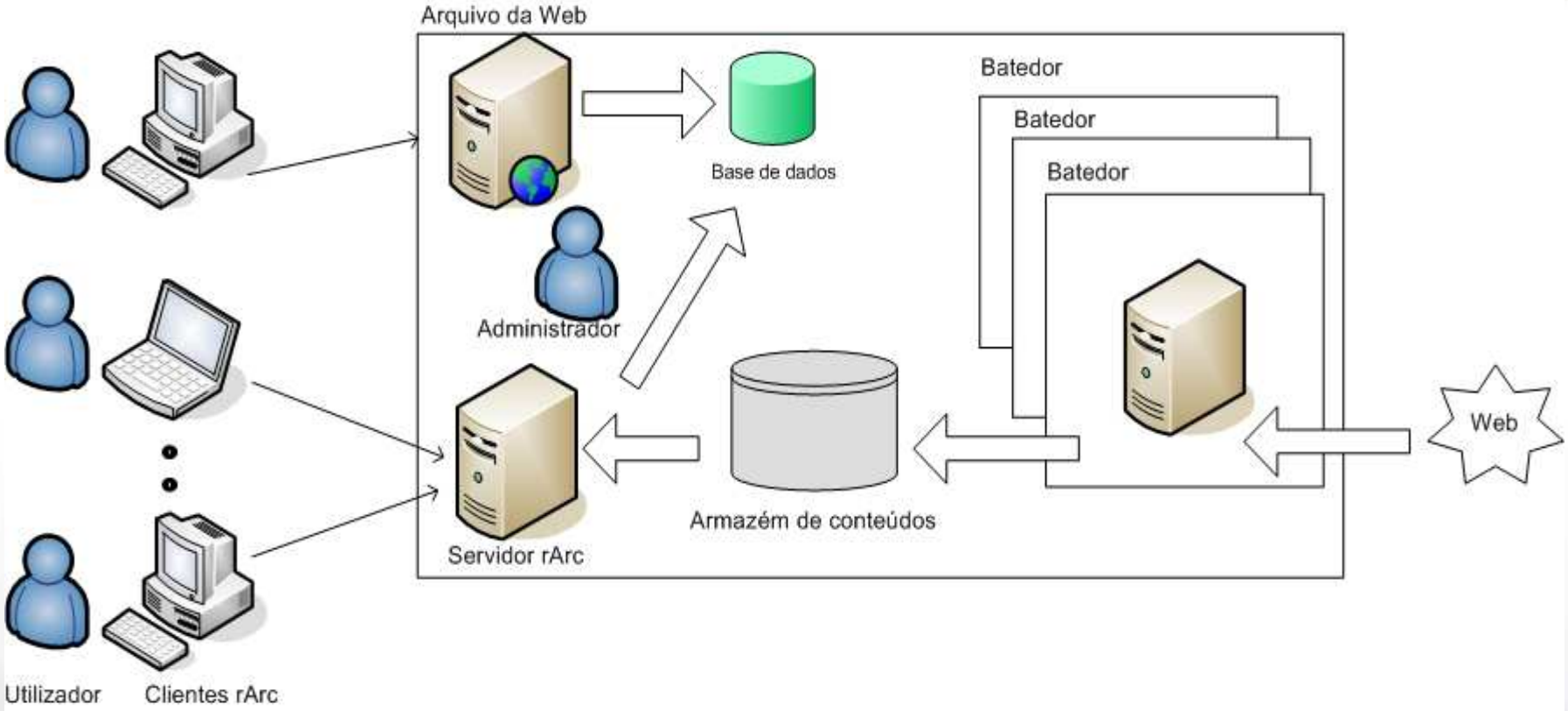
Requisitos de um arquivo da Web

- Disponibilidade
  - Diferentes fusos horários
- Confidencialidade
  - Informação confidencial; *Spam*
- Autenticidade
  - Virus informáticos
- Integridade
  - Problemas físicos dos discos dos computadores dos utilizadores
- Portabilidade
  - Rede mundial de computadores heterogéneos
- Usabilidade
  - Motivação dos utilizadores

Trabalho Relacionado

- Partilha de ficheiros ponto-a-ponto (Napster)
  - Partilha de informação
- Bibliotecas digitais (LOCKSS)
  - Problemas de confidencialidade
- Data grids (SRB)
  - Acesso uniforme à informação
- Computação distribuída voluntária (BOINC)
  - Processamento de informação
- Armazenamento distribuído (OceanStore)
  - Disponível apenas um protótipo

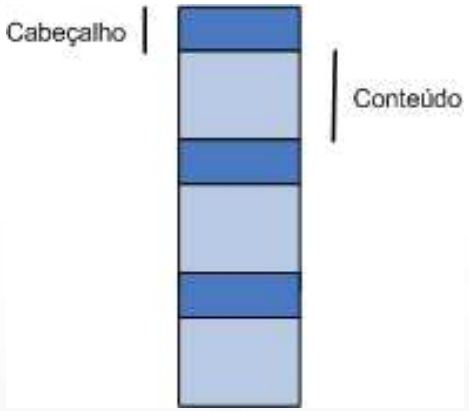
Arquitectura



- **Base de dados sobrevive à perda de informação no armazém de conteúdos**

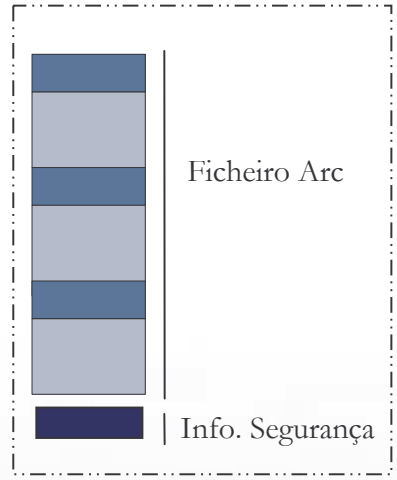
Formato Arc

- Especificação de ficheiro desenvolvido pelo Internet Archive.
- Formato utilizado pelo batedor Heritrix.
- O tamanho definido por omissão no Heritrix é 100 MBytes.
- Warc o formato sucessor do Arc.

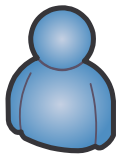


Cápsula

- Permite garantir a confidencialidade e autenticidade
- Uma cápsula é constituída por:
  - Ficheiro Arc
  - Info. Segurança (assinatura simétrica)



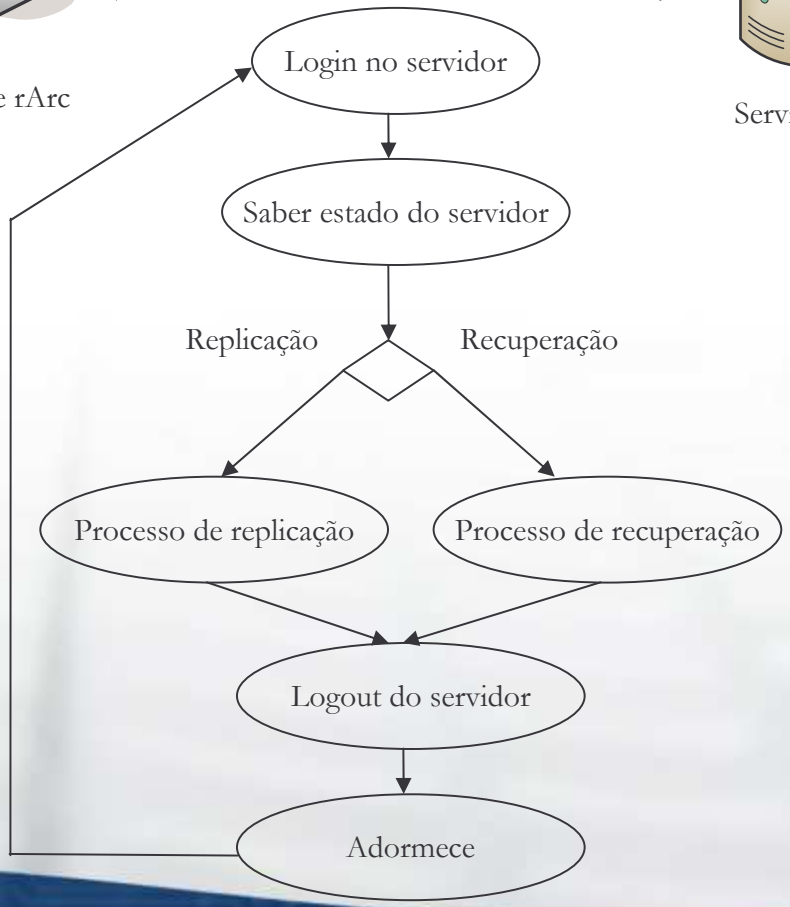
Funcionamento



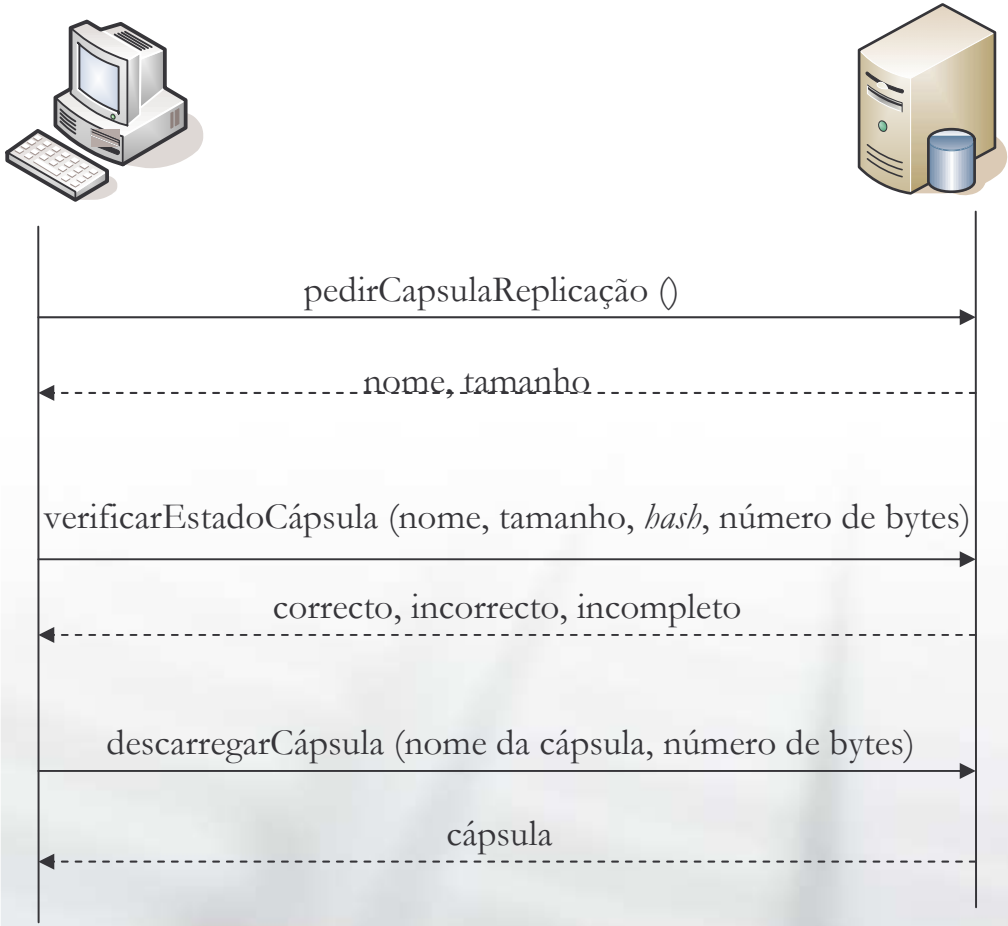
Utilizador

Cliente rArc

Servidor rArc



Processo de replicação



- O servidor escolhe a cápsula com menos tentativas de replicação

Nome de cápsula	Número de vezes eleita para replicação
Cápsula 1	2
Cápsula 2	2
Cápsula 3	2
Cápsula 4	1

- Quando são adicionadas novas cápsulas ao sistema, o servidor vai eleger estas para replicação.

Processo de recuperação

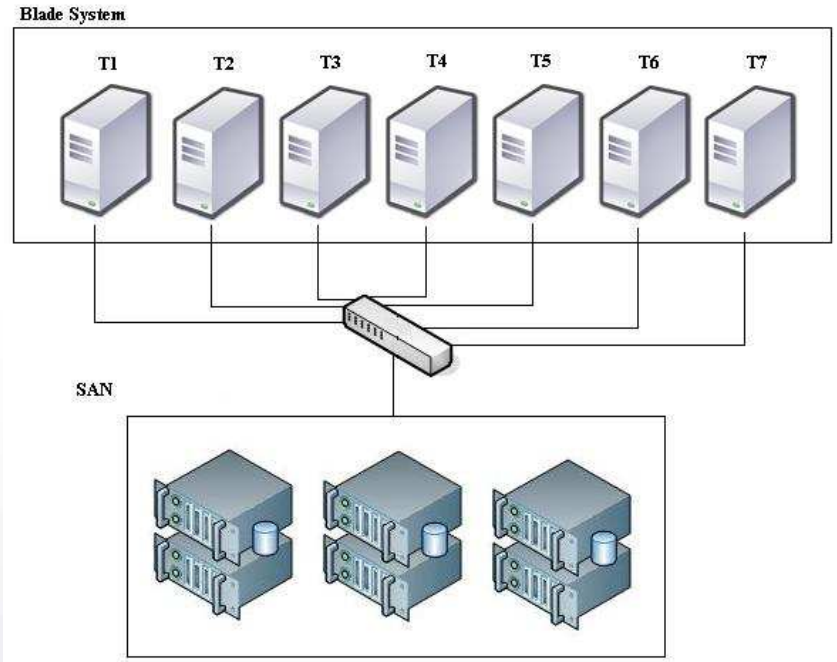
- Quando termina o envio de uma cápsula, o servidor verifica a autenticidade do ficheiro Arc
- Quando uma cápsula é recuperada, as restantes transferências da mesma cápsula são terminadas

Tecnologias

- Ferramentas
  - ArgoUml (análise e desenho)
  - Java (desenvolvimento)
    - Eclipse
  - MySQL (desenvolvimento)
- Ferramentas e tecnologias utilizadas de código aberto
  - Não existe custo de aquisição

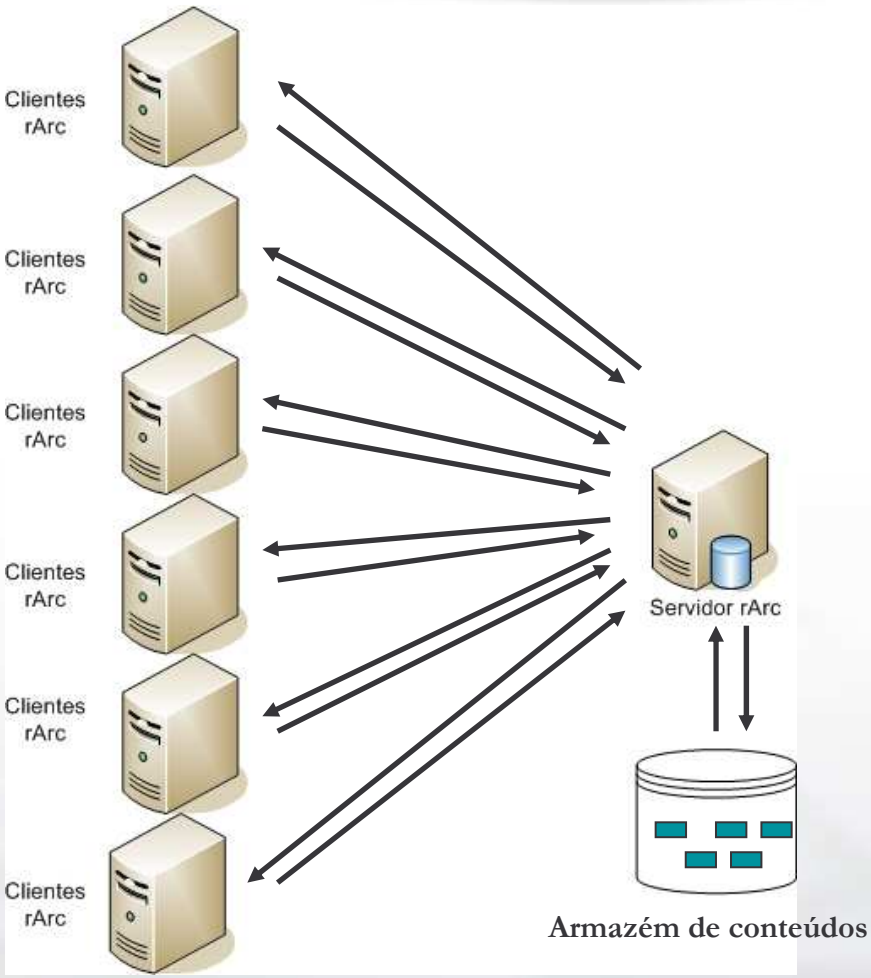
Avaliação

- Objectivo
  - Avaliar a correcção de funcionamento
  - Avaliar o desempenho
- Infra-estrutura do AWP
  - Blade system
    - Intel Xenon Quad-Core (2.33 GHz)
    - 8 GBytes RAM
    - Red Hat Enterprise Linux 5
  - SAN
    - Espaço total: 24,5 TBytes

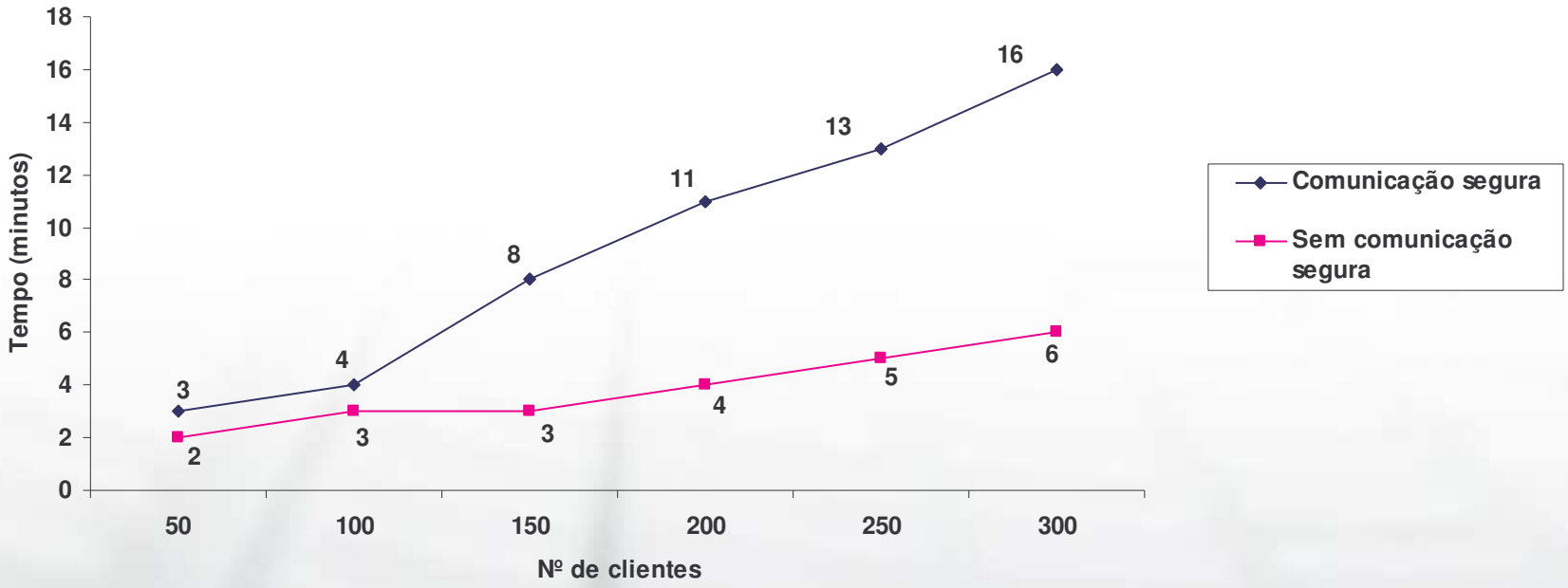


Cenários

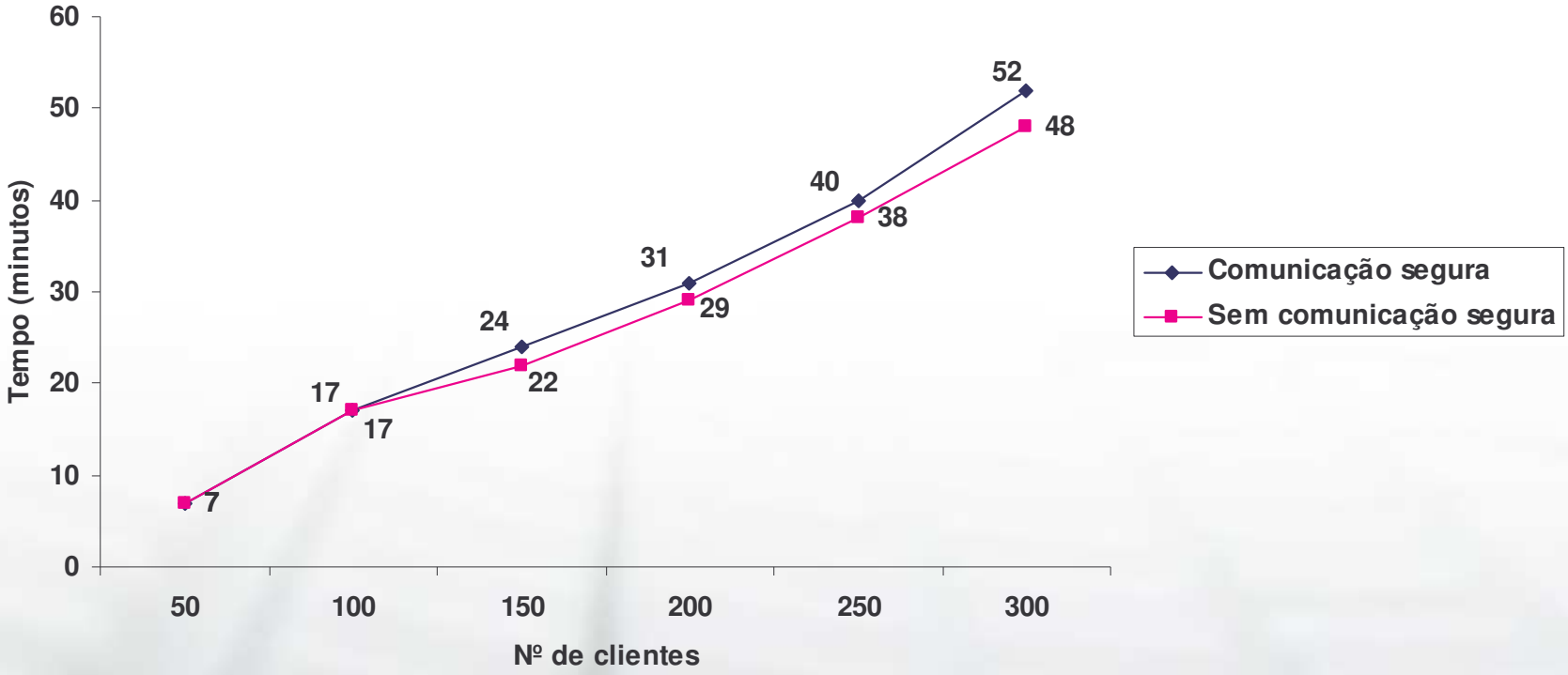
- Dois cenários:
  - Replicação - replicar cápsulas por vários clientes
  - Recuperação - recuperar as cápsulas replicadas pelos clientes
- Medir o tempo total para replicar e recuperar as cápsulas



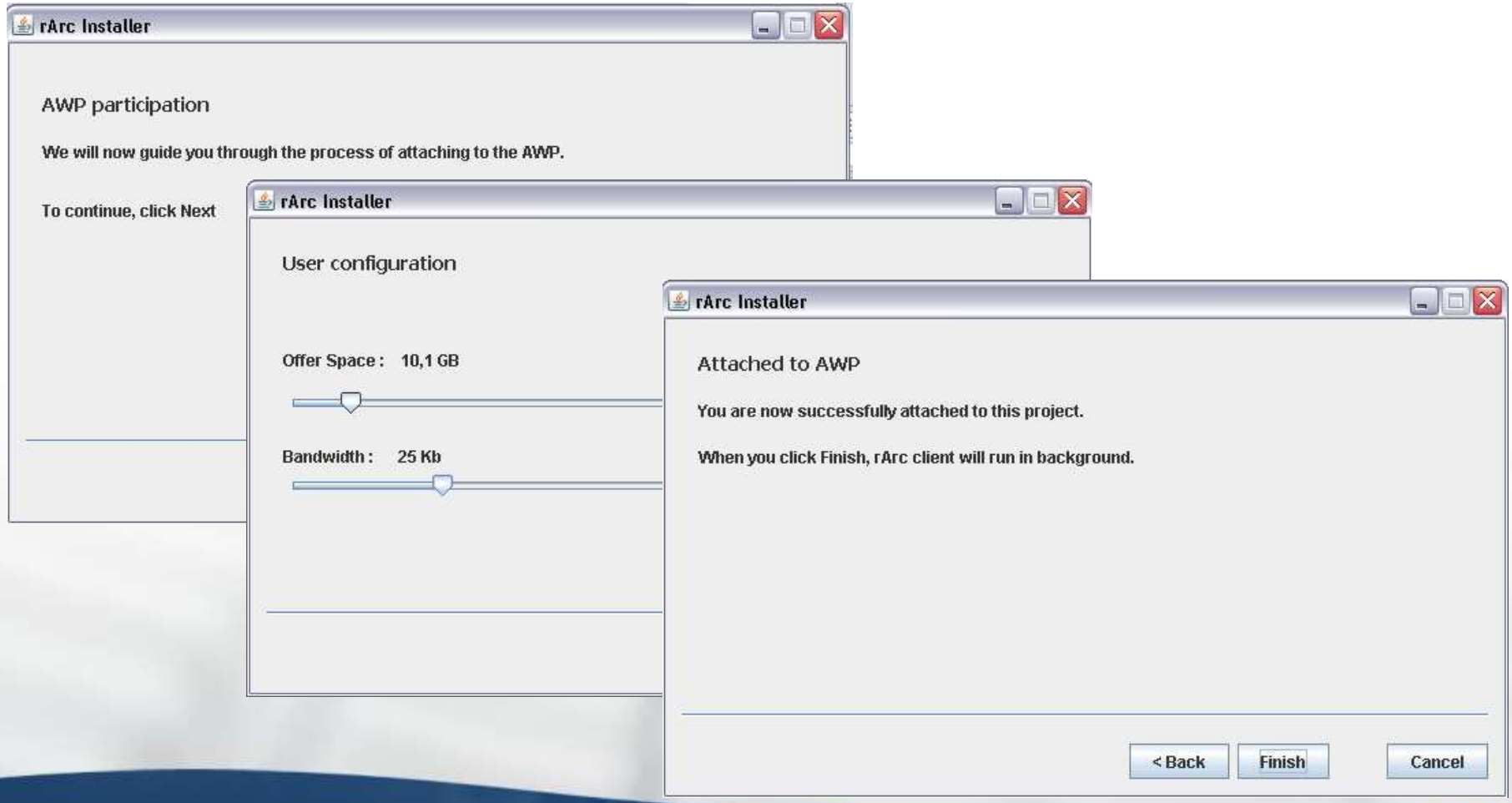
Resultado da replicação



Resultado da recuperação



Trabalho pós-tese



Trabalho Futuro

- Motivação do utilizadores
  - Sítio de apoio
  - Sistema de ranking
- Melhorar o desempenho
- Recuperação parcial
- Comunicação segura apenas na autenticação do cliente rArc

Conclusões

- Análise de requisitos para sistemas relacionados com o arquivo da Web não é trivial
- Não existia um sistema que respondesse aos requisitos dos arquivos da Web.
- Rarc disponível para a comunidade em código aberto
- A comunidade pode contribuir para a preservação dos conteúdos dos arquivos da Web
- Disponível em breve em http://arquivo-web.fccn.pt

Obrigado pela atenção