ChironFS

Nível: Intermediário

Escopo: Apresentação de sistema de arquivos tolerante a falhas com replicação de dados

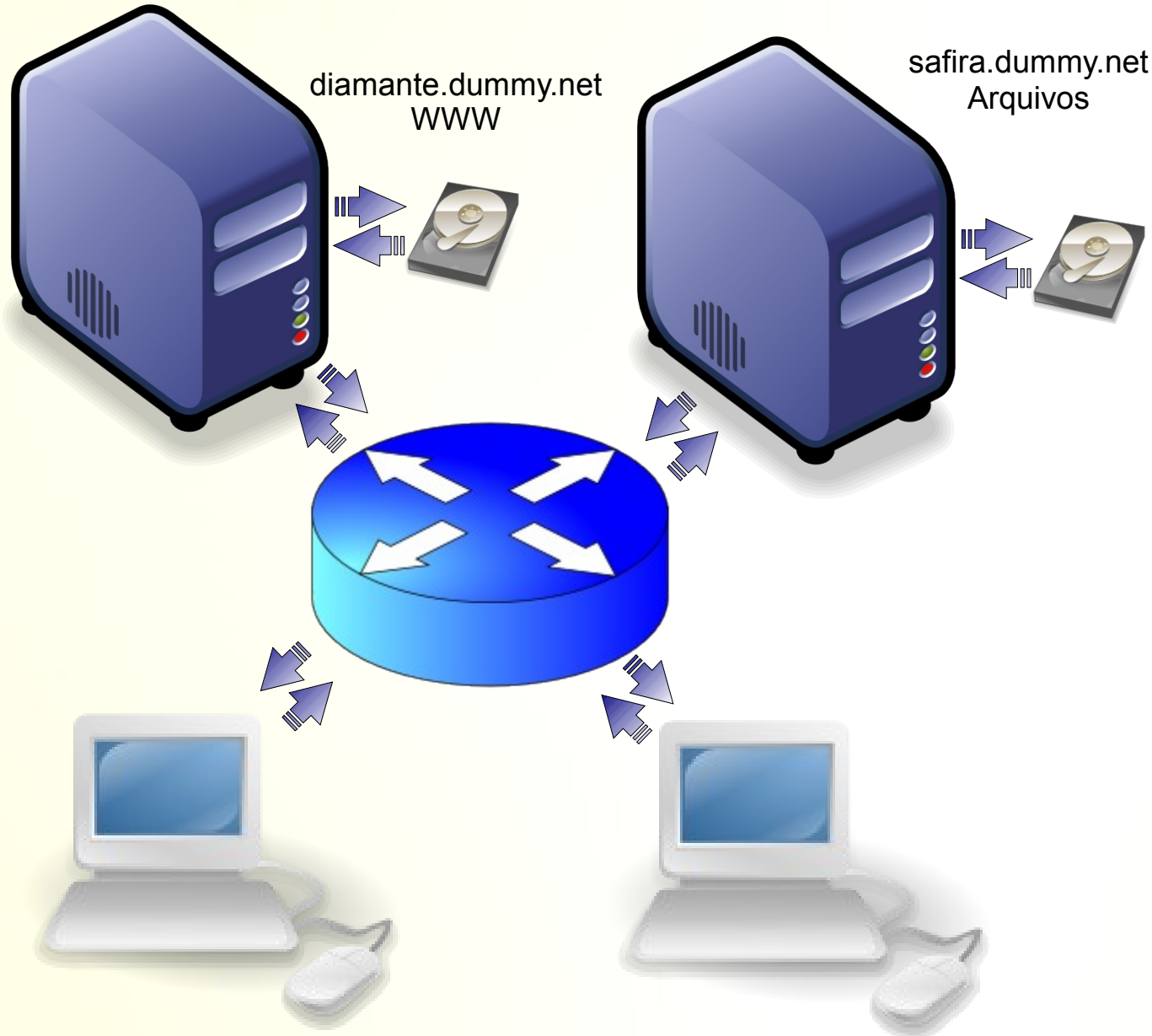
ChironFS

Sistema de Arquivos Tolerante a Falhas com Replicação de Dados

http://www.furquim.org/chironfs/

Luis Otávio de Colla Furquim

Exemplo de Rede Típica



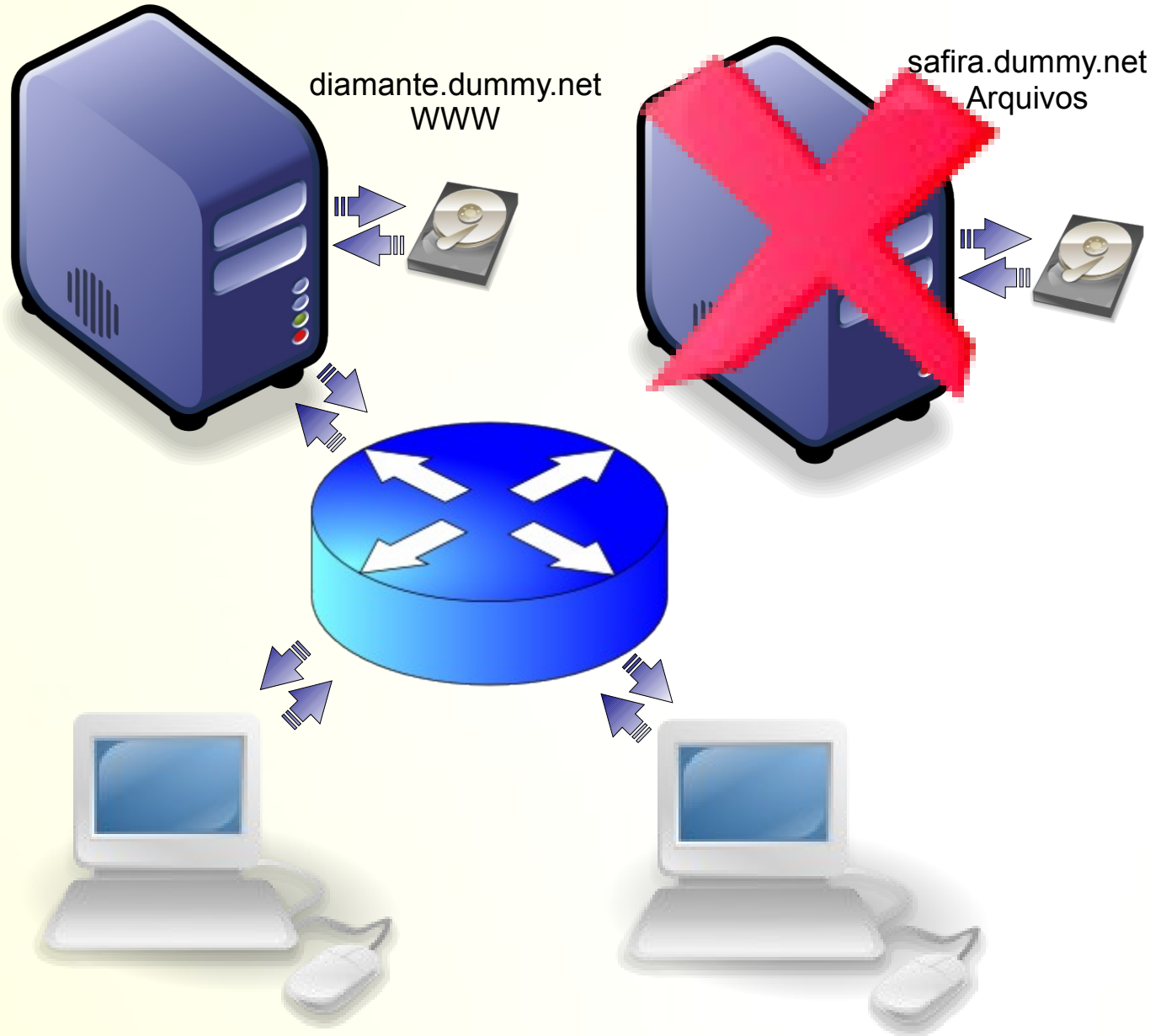
Serviços:

- Web
- Arquivos

Backups:

- todas as noites
- em fita

Falha em uma Rede Típica



Serviço inoperante:

- Compartilhamento de arquivos

Contingência:

- Servidor web acumulará os serviços
- Será restaurado o último backup

Previsão de Reoperacionalização

backup:

- tempo do restore

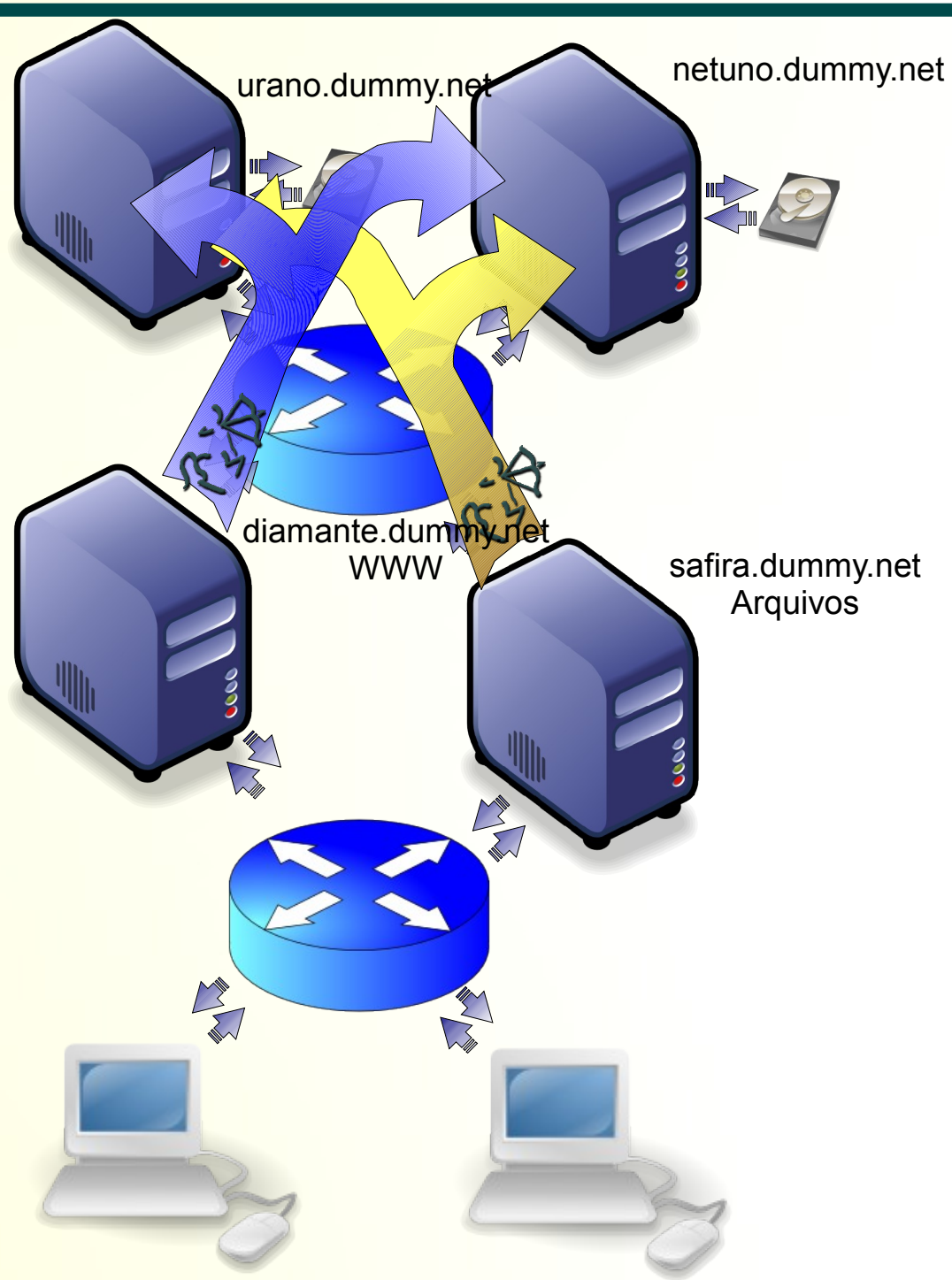
Disco rígido:

- tempo de manutenção do servidor

OU

- nunca (perda dos dados)

Rede com Redundância de Dados



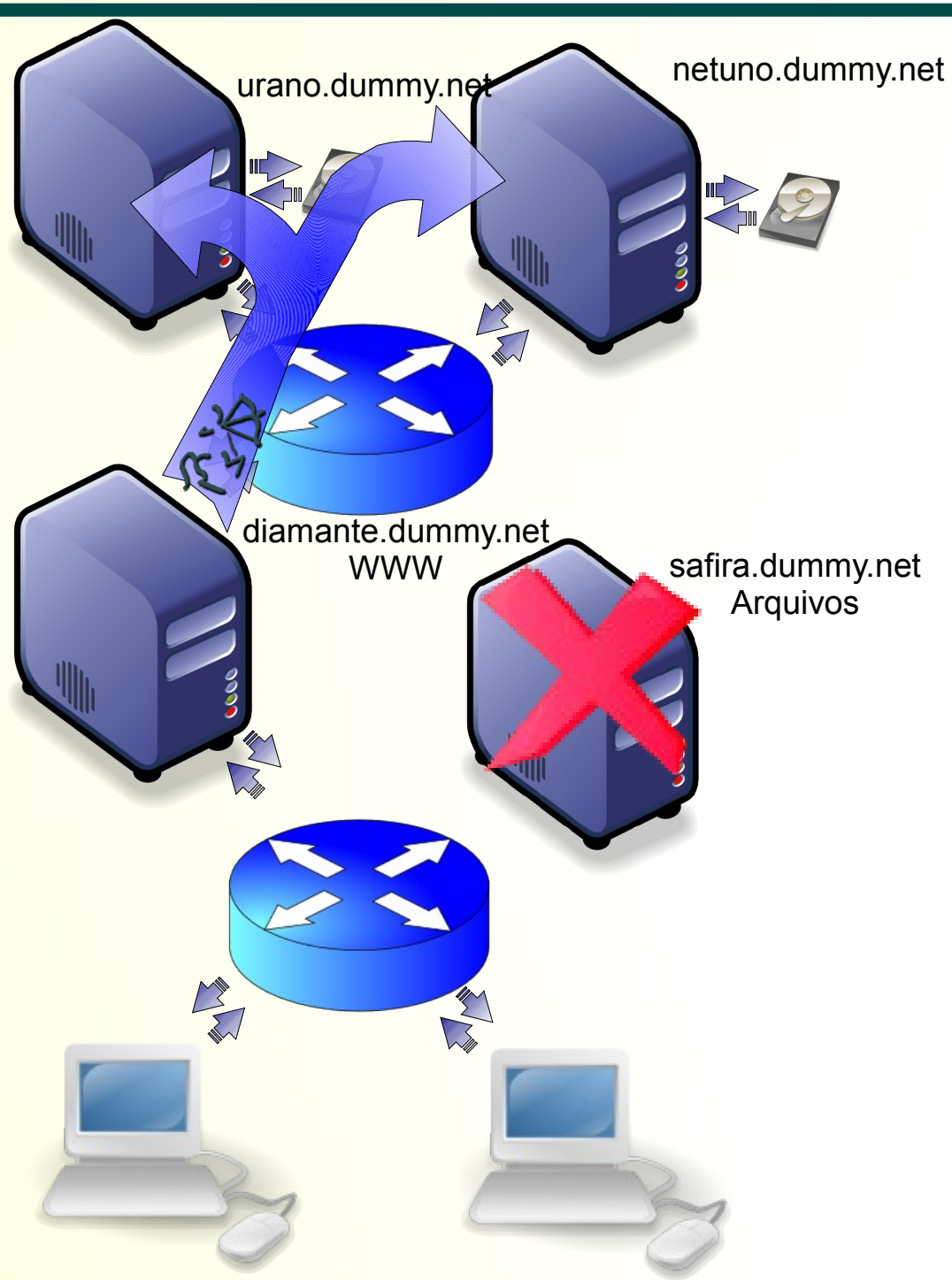
Serviços:

- Web
- Arquivos

Storage:

- redundante
- fisicamente separada dos serviços

Falha de um Serviço



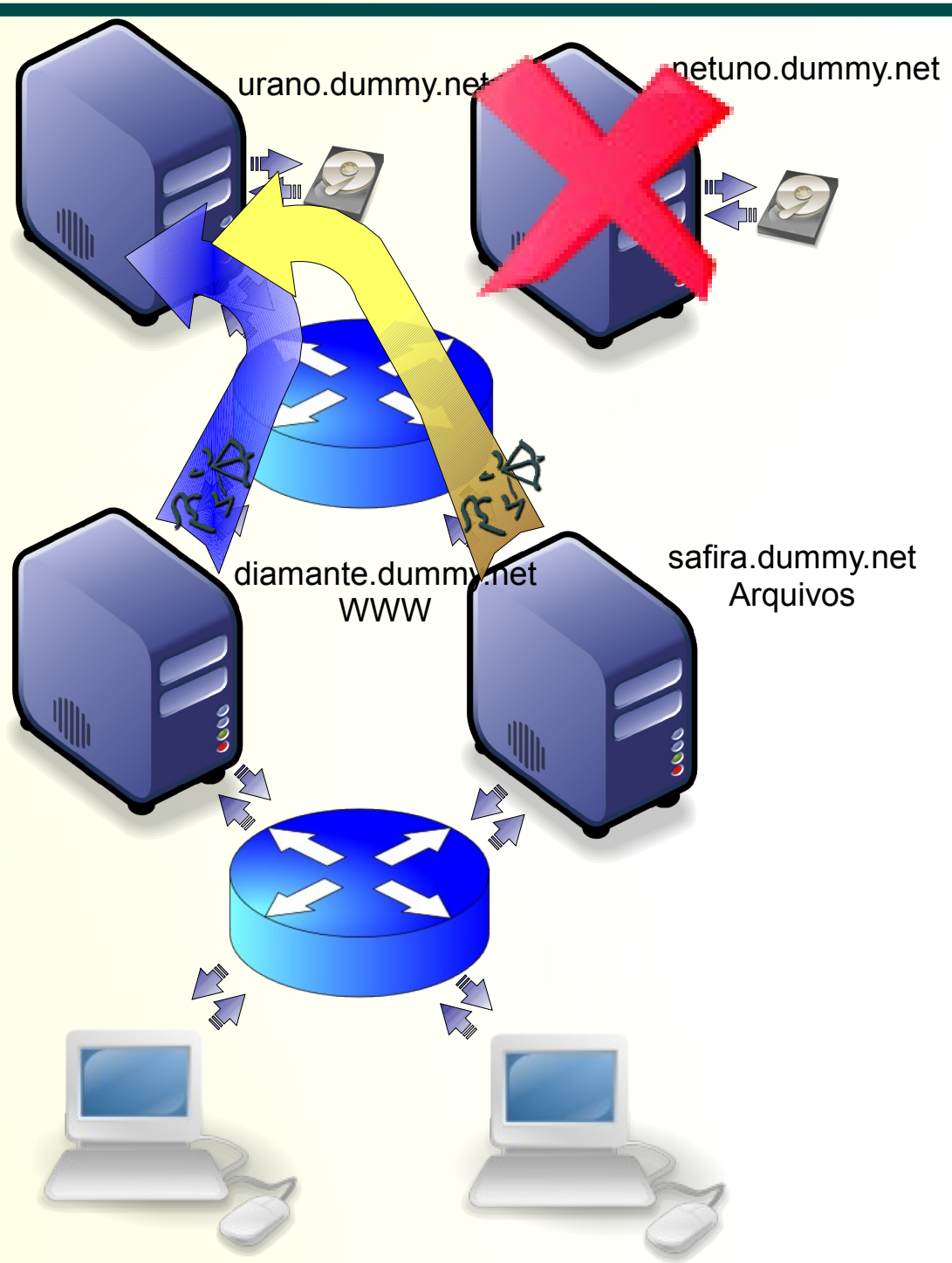
Serviço inoperante:

- Compartilhamento de arquivos

Contingência:

- Servidor web acumulará os serviços
- Dados estão imediatamente disponíveis em qualquer dos nodos do storage

Falha de um Nodo do Storage

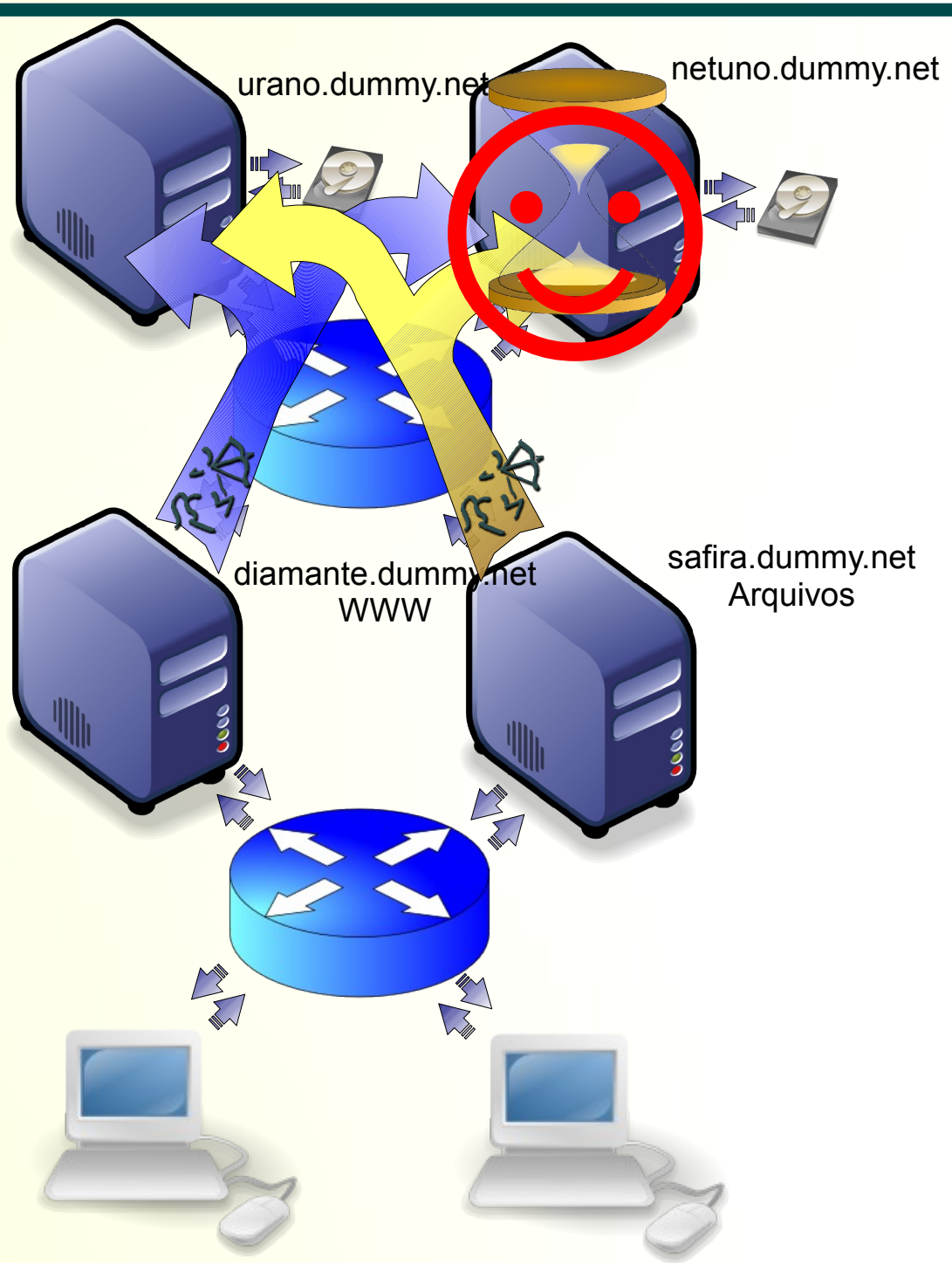


Serviços operantes

Contingência:

- Dados disponíveis no nodo operante

Previsão de Reoperacionalização



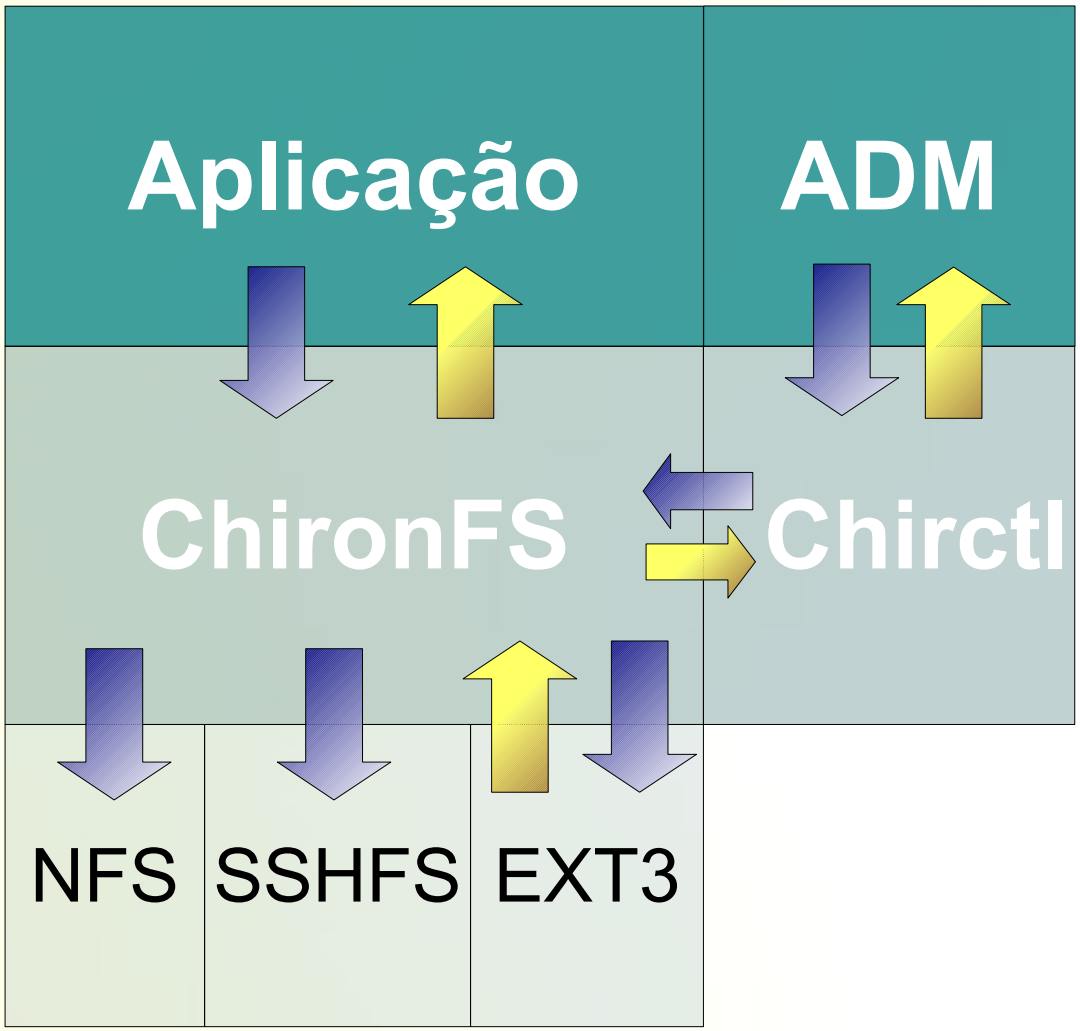
Usuários:

- não percebem a falha

Gerentes de rede:

- reoperacionalização do hardware
- sincronia dos dados

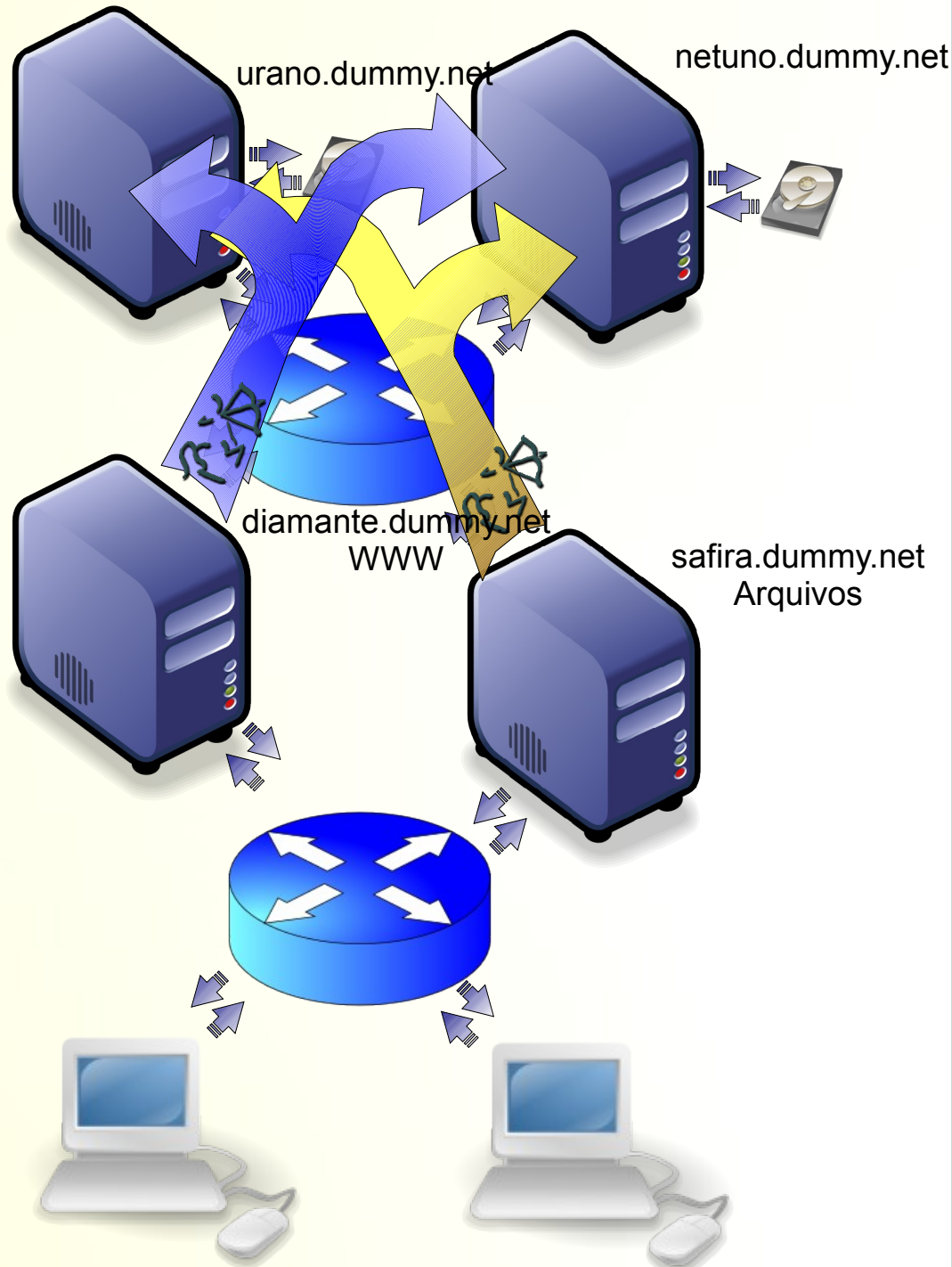
Arquitetura



Filesystem virtual:

- replica filesystems
- sem limite de réplicas
- réplicas em filesystems diferentes
- leitura balanceada
- tolerante a falhas
- protocolo: qualquer um
- autenticação: qualquer uma
- simplicidade de uso
- simplicidade de código
- Filesystem auxiliar para tarefas administrativas

Rede com Redundância de Dados



mkdir /real1 /real2 /virtual /ctl

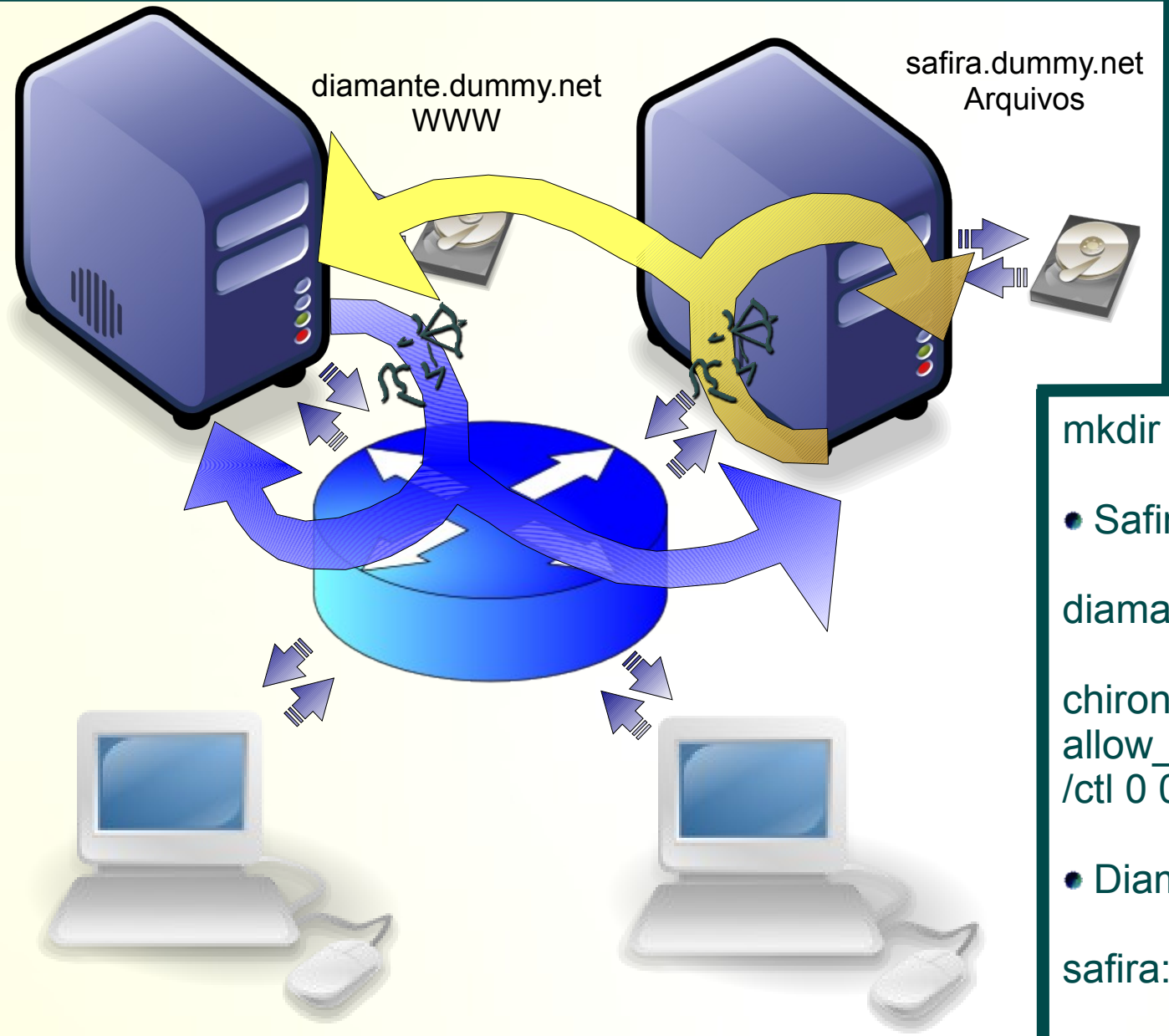
/etc/fstab:

urano:/data /real1 nfs auto 0 0

netuno:/data /real2 nfs auto 0 0

chironfs#/real1=/real2 /virtual fuse allow_other,log=/var/log/chironfs.log,ctl=/ctl 0 0

Redundância sem Storage



- cópia local
- cópia remota em outro servidor de aplicação

mkdir /real1 /real2 /virtual /ctl

- Safira:/etc/fstab

diamante:/real1 /real1 nfs auto 0 0

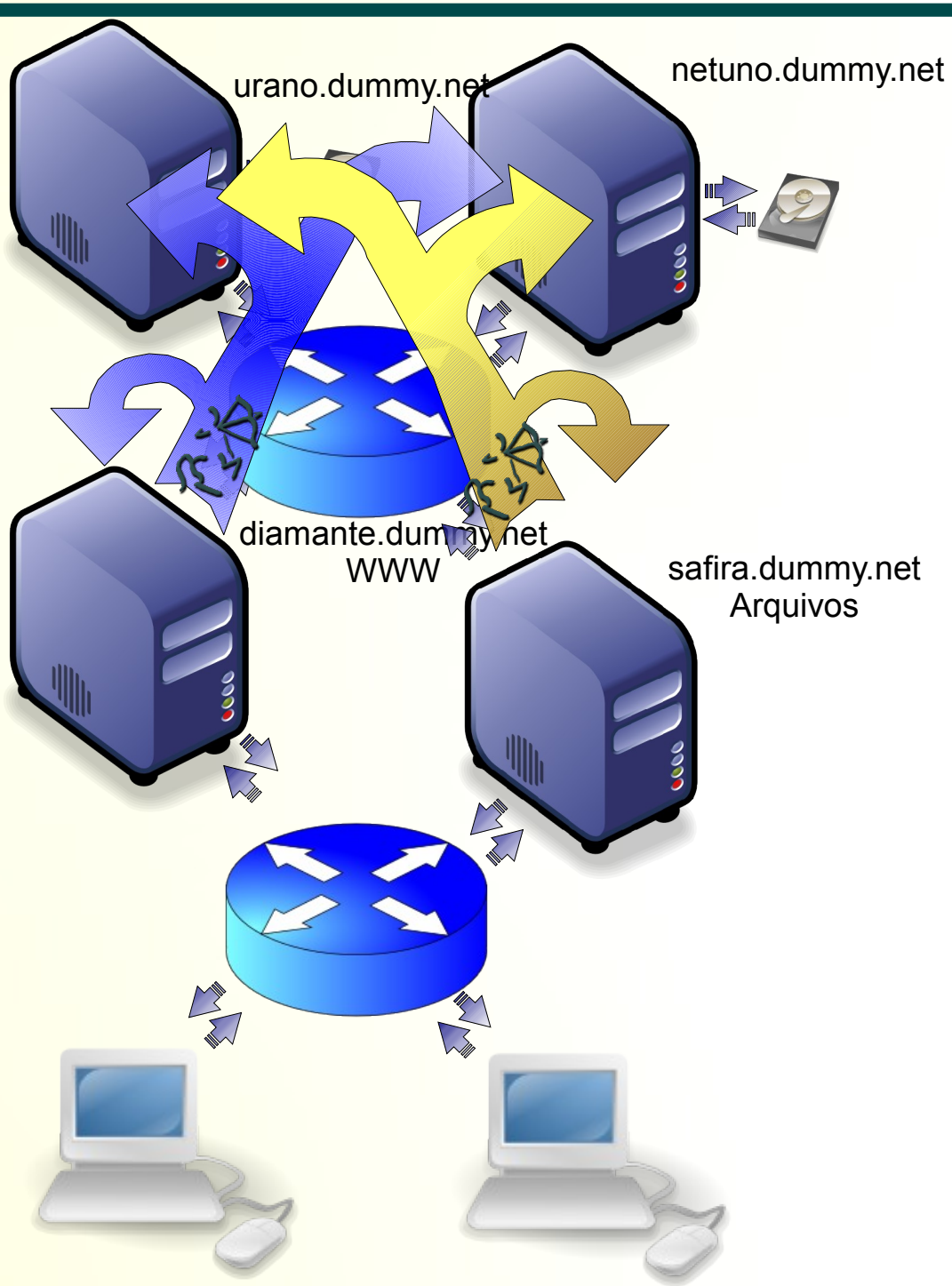
chironfs#/real2=:/real1 /virtual fuse allow_other,log=/var/log/chironfs.log,ctl=/ctl 0 0

- Diamante:/etc/fstab

safira:/real2 /real2 nfs auto 0 0

chironfs#/real1=:/real2 /virtual fuse allow_other,log=/var/log/chironfs.log,ctl=/ctl 0 0

Redundância Mista: Storage e Local



- cópia local
- cópias nos nodos do storage

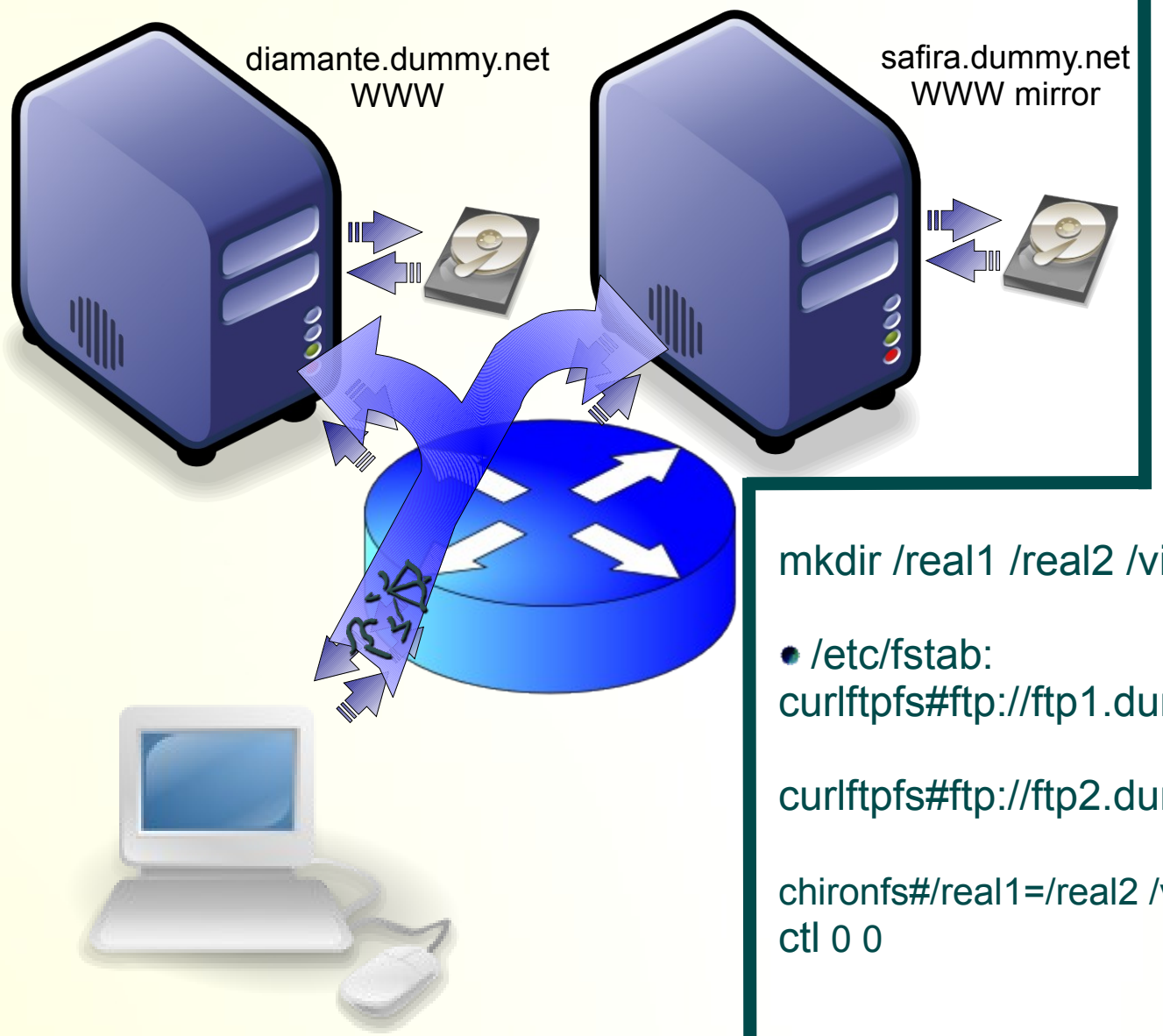
mkdir /real1 /real2 /real3 /virtual /ctl

/etc/fstab:
urano:/data /real1 nfs auto 0 0

netuno:/data /real2 nfs auto 0 0

chironfs#/real3=:/real2=:/real1 /virtual fuse allow_other,log=/var/log/chironfs.log,ctl=/ctl 0 0

Download Balanceado



- Estação distribui a carga entre os mirrors
- Combinação com sistemas de arquivos de FTP e HTTP

mkdir /real1 /real2 /virtual /ctl

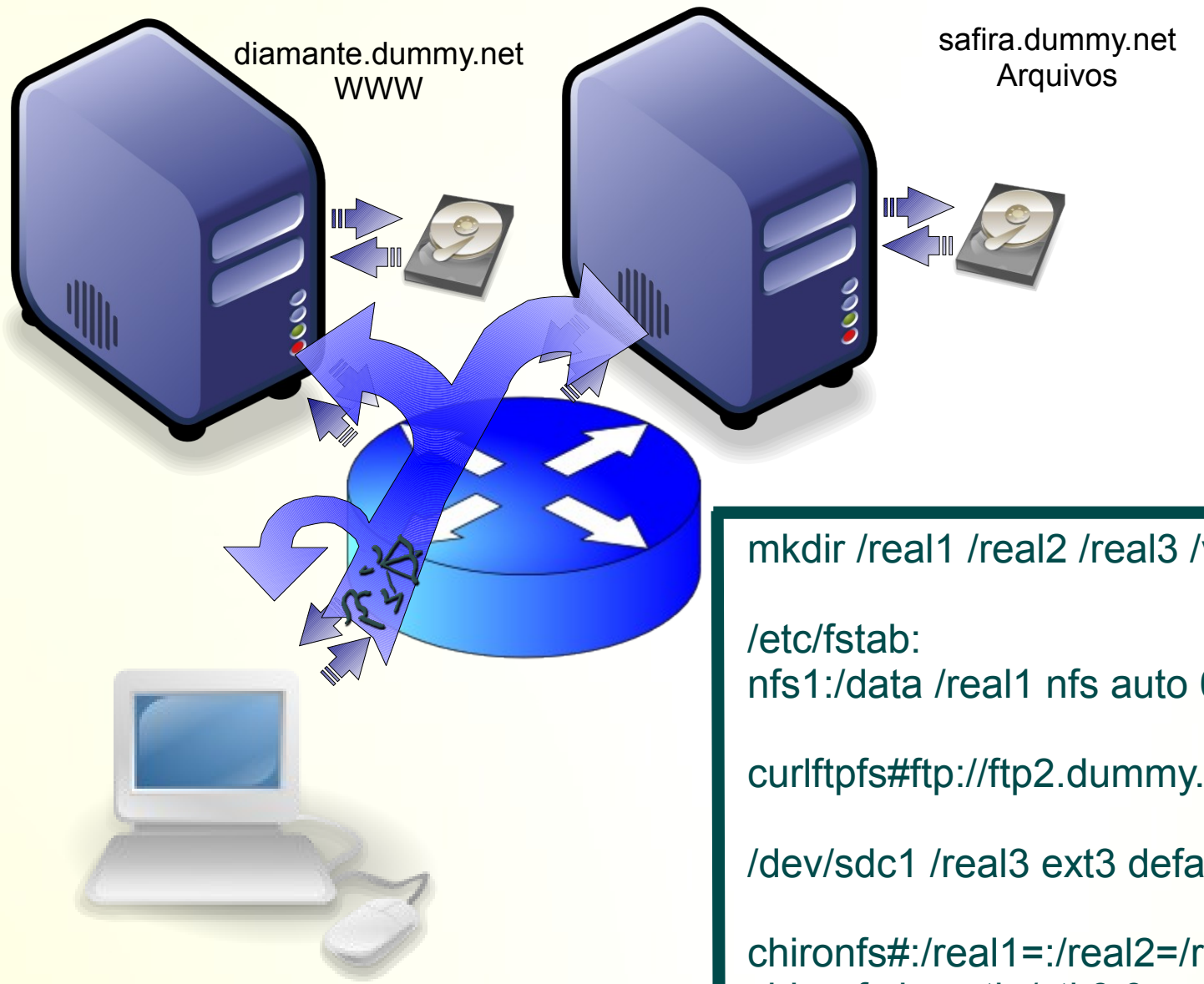
- /etc/fstab:

curlftpfs#ftp://ftp1.dummy.net/ /real1 fuse auto 0 0

curlftpfs#ftp://ftp2.dummy.net/ /real2 fuse auto 0 0

chironfs#/real1=/real2 /virtual fuse log=/var/log/chironfs.log,ctl=/ctl 0 0

Backup de Desktop em Rede



- cópia local
- cópia em servidores de arquivo (LAN/WAN)
- cache local (v1.2)

```
mkdir /real1 /real2 /real3 /virtual /ctl

/etc/fstab:
nfs1:/data /real1 nfs auto 0 0

curlftpfs#ftp://ftp2.dummy.net/ /real2 fuse auto 0 0

/dev/sdc1 /real3 ext3 defaults 0 1

chironfs#:/real1=:/real2=/real3 /virtual fuse log=/var/log/
chironfs.log,ctl=/ctl 0 0
```

Deficiências

- Ressincronia das réplicas falhas a cargo do administrador
- ~~Reintegração da réplica falha via remontagem do filesystem~~
- Gravação síncrona das réplicas:

Tempo	1	2	3	4	5	6	7	8	9	10	11
/real1	■	■	■								
/real2				■	■	■					
/real3							■	■	■		
/real1			■	■	■						
/real2						■	■	■			
/real3									■	■	■

Solução: versão 1.2

- Interface de controle semelhante a /proc:
  - Maior flexibilidade na ressincronia de réplicas falhas
  - Acesso concorrente ✓
  - Reintegração de réplicas sem remontagem do filesystem ✓
  - Realização de consultas ✓
- Gravação assíncrona das réplicas:

Tempo	1	2	3	4	5	6	7	8	9	10	11
/real1	■	■	■								
/real2		■	■	■							
/real3			■	■	■						
/real1			■	■	■						
/real2				■	■	■					
/real3					■	■	■				

Outras versões

- Suporte a atributos extendidos
- Sistema de plugins para algoritmo de baleceamento de leitura
- Estatísticas de acesso disponíveis ao plugin de balanceamento
- Tamanho/espaço livre com réplicas de diferentes tamanhos
  - Reportar o menor tamanho: evita desabilitar réplica menor
  - Reportar o maior tamanho: réplica menor = canário