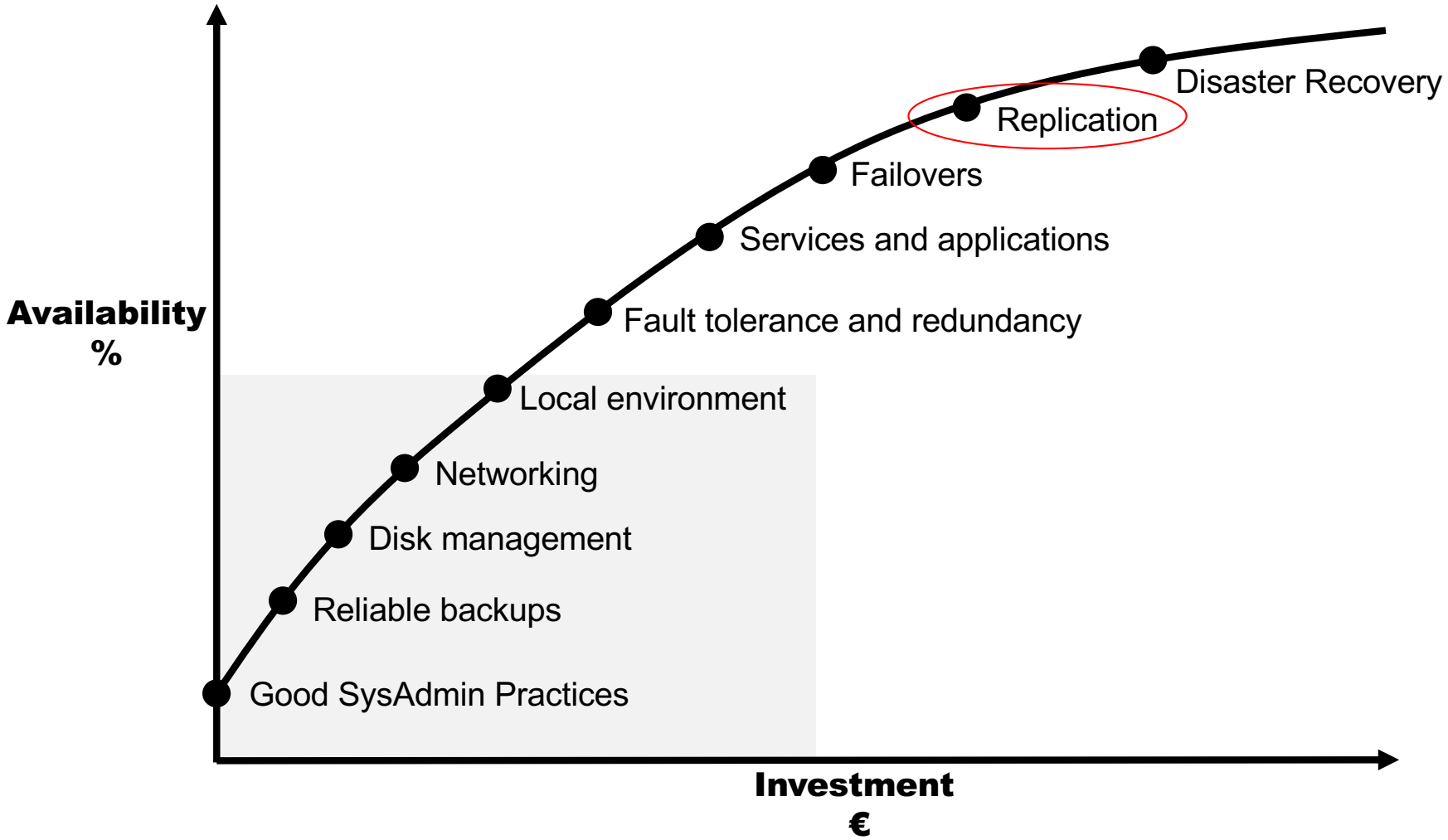
Replicação de dados

1. Tipos de replicação de dados
2. Tecnologias e protocolos mais utilizados
3. RAID
4. Noção de SAN
5. SCSI
6. iSCSI
7. Fiber-Channel

mario.antunes@ipleiria.pt

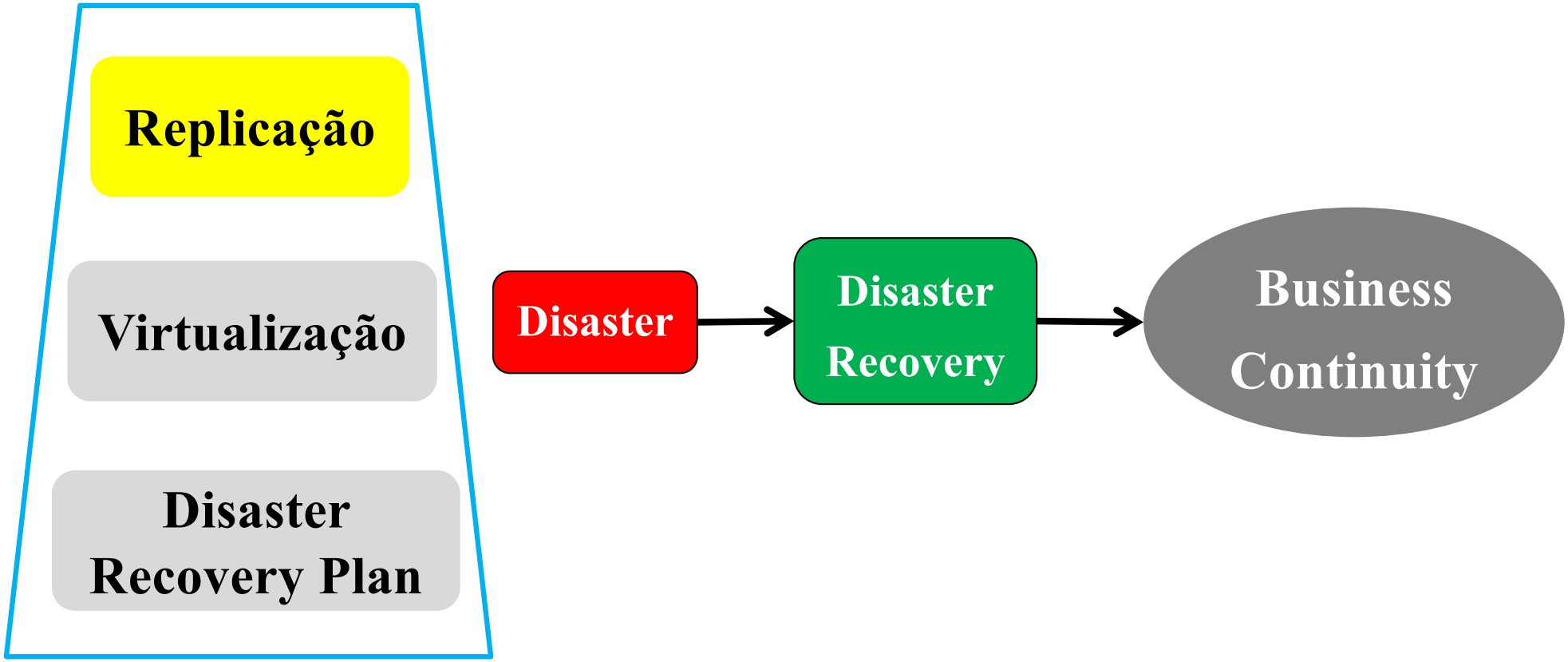
Enquadramento



Adapted by Marcus E, Stern H., *"Blueprints for high availability";* 2003; *Wiley;* ISBN: 0471430269;

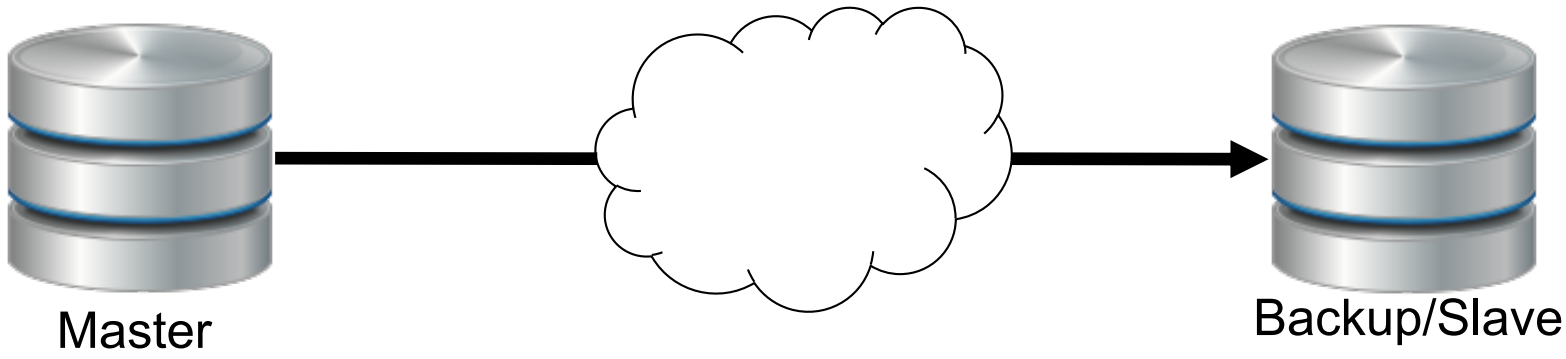
mario.antunes@ipleiria.pt

Enquadramento



mario.antunes@ipleiria.pt

Enquadramento



REPLICAÇÃO ≠ MIRRORING ≠ PARTILHA

- A replicação trata os conjuntos de discos separadamente
- Os discos de cada site podem ter uma configuração em RAID
- Replicação assegura a existência de duas cópias consistentes
- Numa estratégia de DR, cópias estão fisicamente distantes.

mario.antunes@ipleiria.pt

Motivação

- Operações de disaster recovery. Recuperação dos dados do negócio após falha do site principal
- Dados no site de backup podem ser usados para outras tarefas, sem comprometer site principal (p.e. reports, data-mining, ...)
- Nalgumas situações, site de backup pode ser usado para estratégias de *failover*.

mario.antunes@ipleiria.pt

Tipos de replicação

Latency-based

- Síncrona
- Assíncrona
- Periódica

Initiator-based

- Hardware
- Software
- Filesystem
- Aplicação (p.e. DB)
- Transações

mario.antunes@ipleiria.pt

Tipos de replicação – latency-based - síncrona

- Cópia simultânea entre os nós master e slave
- Latência entre a cópia dos dados pela rede e a respetiva confirmação
- Distância pode variar de acordo com a tecnologia usada
- Garante sincronismo entre as cópias dos site principal e do de DR
- Consistência garantida pela atomicidade das operações.
- Exemplo: BD distribuídas, via cloud (Google, Amazon, ...)

Solução que garante o mínimo de perda em caso de desastre

mario.antunes@ipleiria.pt

Tipos de replicação – latency-based - assíncrona

- Dados são guardados localmente no servidor master
- Cópia para o destino é feito de acordo com condições definidas: largura de banda, carga do servidor, etc...
- Diminuição da periodicidade de cópia melhora atualização do slave

Perda de dados como compromisso do tempo de latência

© Mário Antunes
mario.antunes@ipleiria.pt

Tipos de replicação – latency-based - periódica

- Backup do master é realizado periodicamente e enviado pela rede para o slave.
- O modo de envio dos dados pela rede é manual.
- A periodicidade de realização do backup é definida manualmente

Perda de dados como compromisso da periodicidade da cópia

mario.antunes@ipleiria.pt

Tipos de replicação – exemplo da Google

Americas

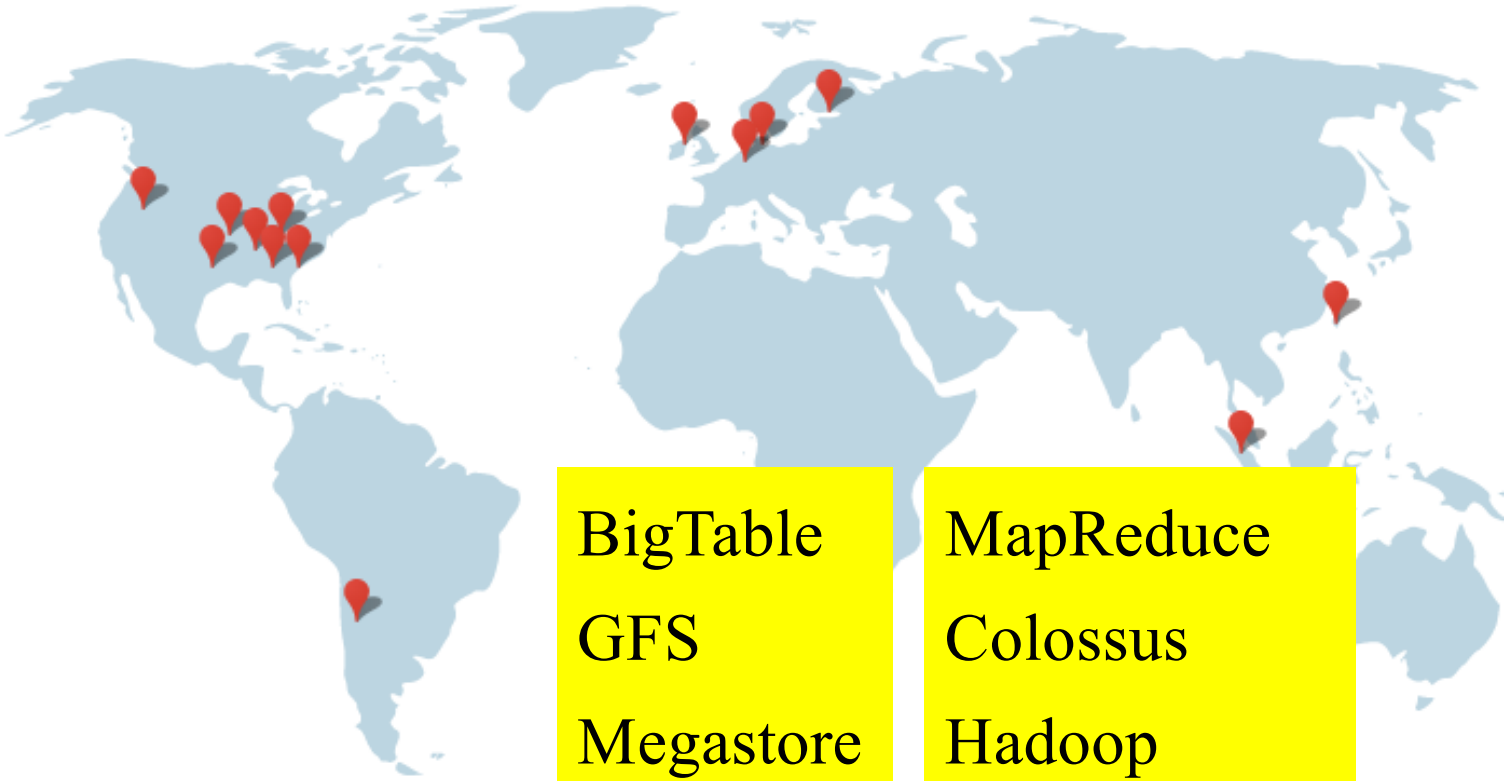
Berkeley County, South Carolina
Council Bluffs, Iowa
Douglas County, Georgia
Quilicura, Chile
Jackson County, Alabama
Mayes County, Oklahoma
Lenoir, North Carolina
The Dalles, Oregon

Asia

Changhua County, Taiwan
Singapore

Europe

Hamina, Finland
St Ghislain, Belgium
Dublin, Ireland
Eemshaven, Netherlands



BigTable
GFS
Megastore

MapReduce
Colossus
Hadoop
Borg

http://www.google.com/about/datacenters/inside/locations/
http://www.datacentermap.com

mario.antunes@ipleiria.pt

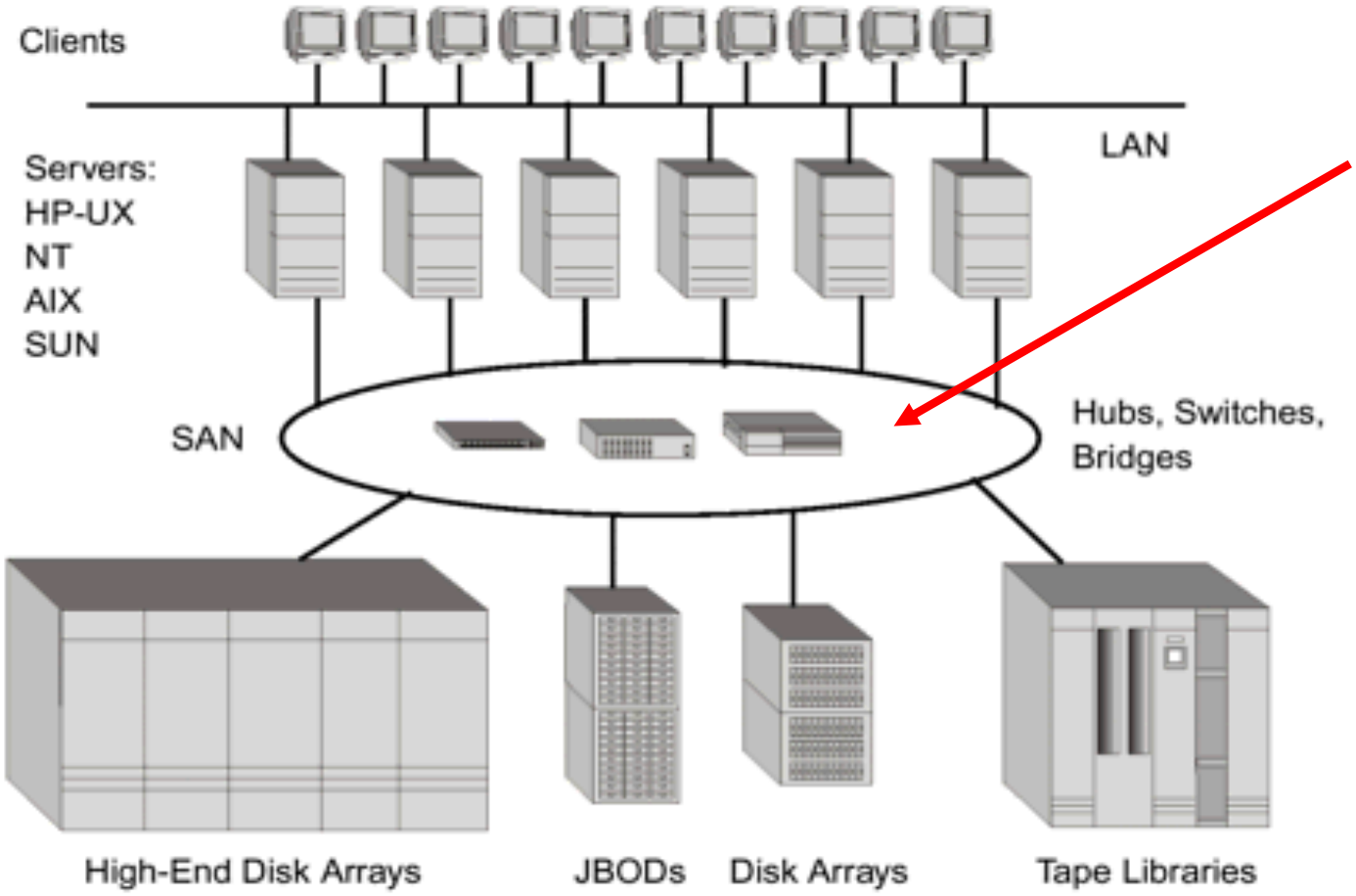
Tipos de replicação – failover remoto

- Deteção automática da falha no site principal
- Promoção automática do site de backup a principal
- Disponibilizar automaticamente os recursos principais
- Arrancar com as aplicações críticas no site de DR

- Clusters remotos de HA com monitorização dedicada
- Aplicam-se os conceitos tradicionais de clusters de HA

mario.antunes@ipleiria.pt

Noção de SAN



SCSI

FiberChannel

iSCSI

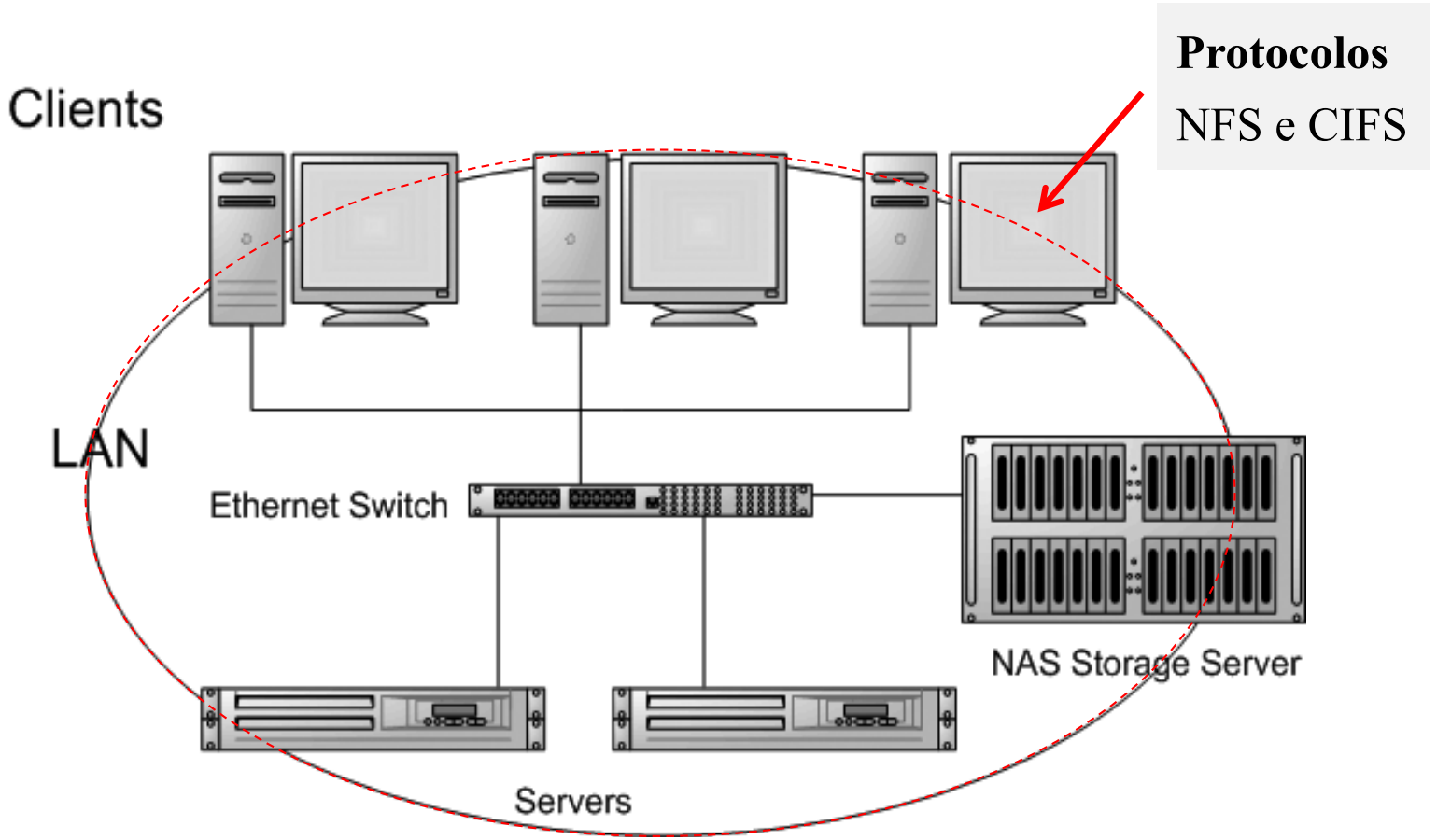
(...)

Noção de SAN

- Acesso partilhado ás unidades de storage (discos, tapes, ...)
- Acesso ao storage numa perspetiva *block-based*
- Acesso transparente para os utilizadores
- Facilita a gestão de storage e é fator chave na replicação de dados e em *disaster recovery*
- Protocolos mais utilizados: FiberChannel (FC) e SCSI.
- Implementação "over IP" (iFCP e iSCSI) permite topologias de área alargada.

mario.antunes@ipleiria.pt

Noção de NAS



http://programming4us.eu

mario.antunes@ipleiria.pt

Noção de NAS

- Centralização do storage num servidor dedicado, com RAID e outras funções de redundância disponibilidade.
- Acesso pela rede local, essencialmente sobre a rede TCP/IP
- Acesso ao storage numa perspetiva *file-based*.
- Acesso transparente para os utilizadores por protocolos NFS e CIFS
- Servidores dedicados para NAS: FreeNAS, NAS4Free, ...

mario.antunes@ipleiria.pt

Sistema RAID

- **R**edundant **A**rray of **I**ndependent/Inexpensive **D**rives
- Os dados são replicados por vários discos
- RAID
  - Hardware: transparente ao sistema operativo
  - Software: implementado ao nível do sistema operativo
- Conceitos chaves
  - Replicação (*mirroring)*
  - Particionamento dos dados por vários discos (*stripping*)

mario.antunes@ipleiria.pt

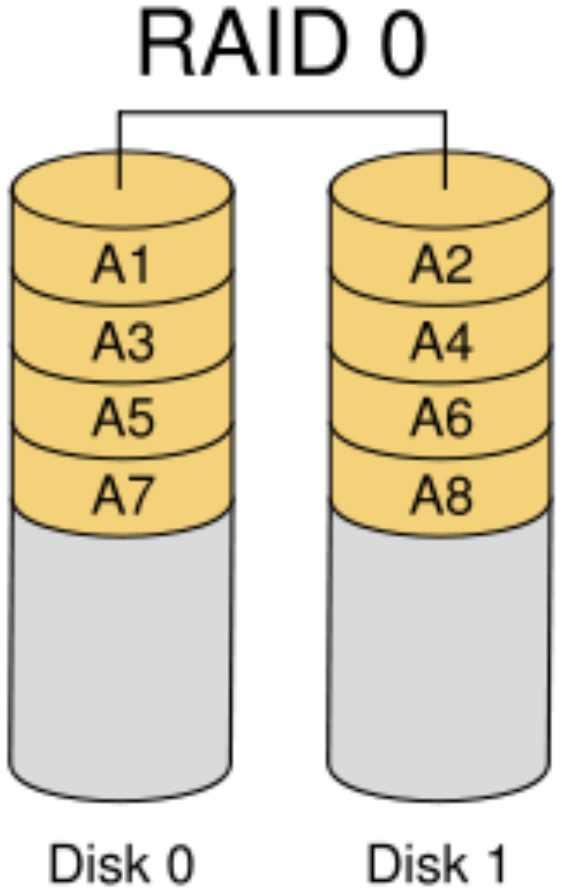
Níveis de RAID

RAID 0 (*striping*)

- Cada ficheiro é particionado
- Respectivos blocos (e.g. 1,A2,A3,...) guardados em cada um dos discos

- Aumenta o desempenho - A leitura de um ficheiro pode ocorrer em paralelo (A1 e A2 podem ser lidos ao mesmo tempo, dado que estão em discos diferentes)

- Não oferece redundância adicional
  Se **um** disco falhar, os dados ficam perdidos...

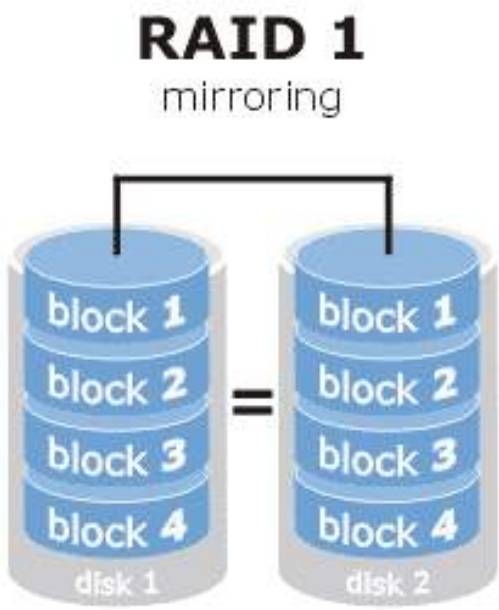


mario.antunes@ipleiria.pt

Níveis de RAID

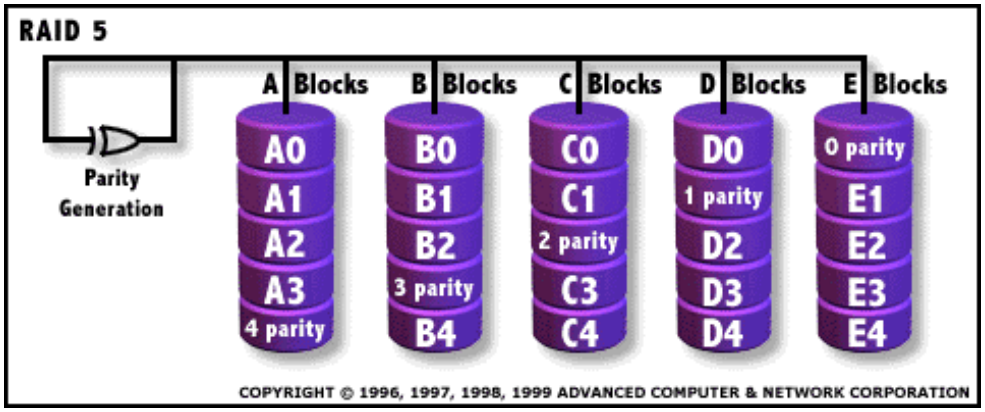
- **RAID 1 (*mirroring*)**

- O disco 1 é uma réplica do disco 0
- Não há melhoria do desempenho
- Há melhoria da tolerância a falhas



Níveis de RAID

- **RAID 5**
  - Mínimo 3 discos
  - Paridade distribuída



  - Para cada bloco de dados existe um bloco de paridade **noutro** disco

  - Tolera a falha de **um** disco

    Se um disco avariar, o sistema mantém-se operacional. O disco em falta pode ser recuperado através da paridade
  - É contudo necessário recuperar o sistema (sistema está vulnerável à falha de um segundo disco)

© Mário Antunes
mario.antunes@ipleiria.pt

Ainda sobre o RAID

Um sistema RAID só protege de falha(s) de hardware:

→Não protege de acidentes provocados por humanos/software, ...

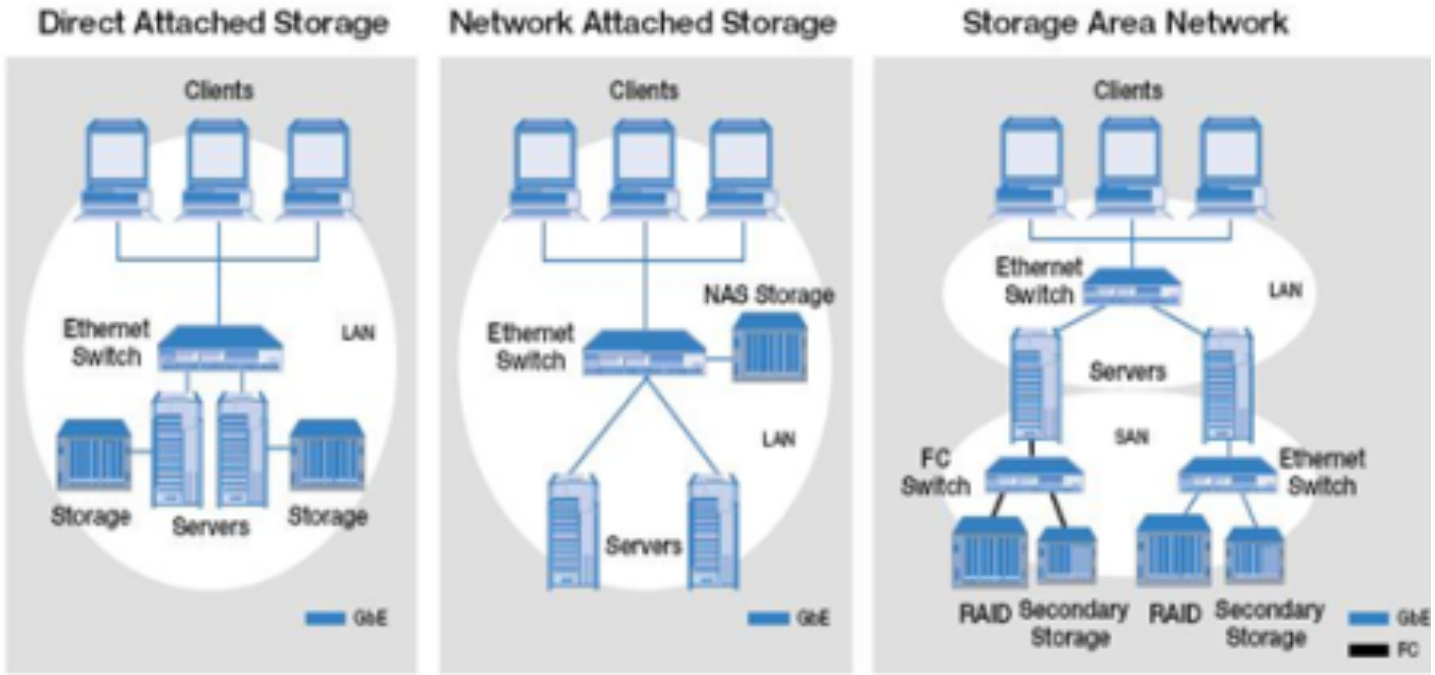
Portanto há sempre necessidade de complementar RAID com sistemas de salvaguarda da informação

→ O RAID aumenta a disponibilidade e induz alguma tolerância a falhas ...

→ ...mas NÃO substitui os backups!

© Mário Antunes
mario.antunes@ipleiria.pt

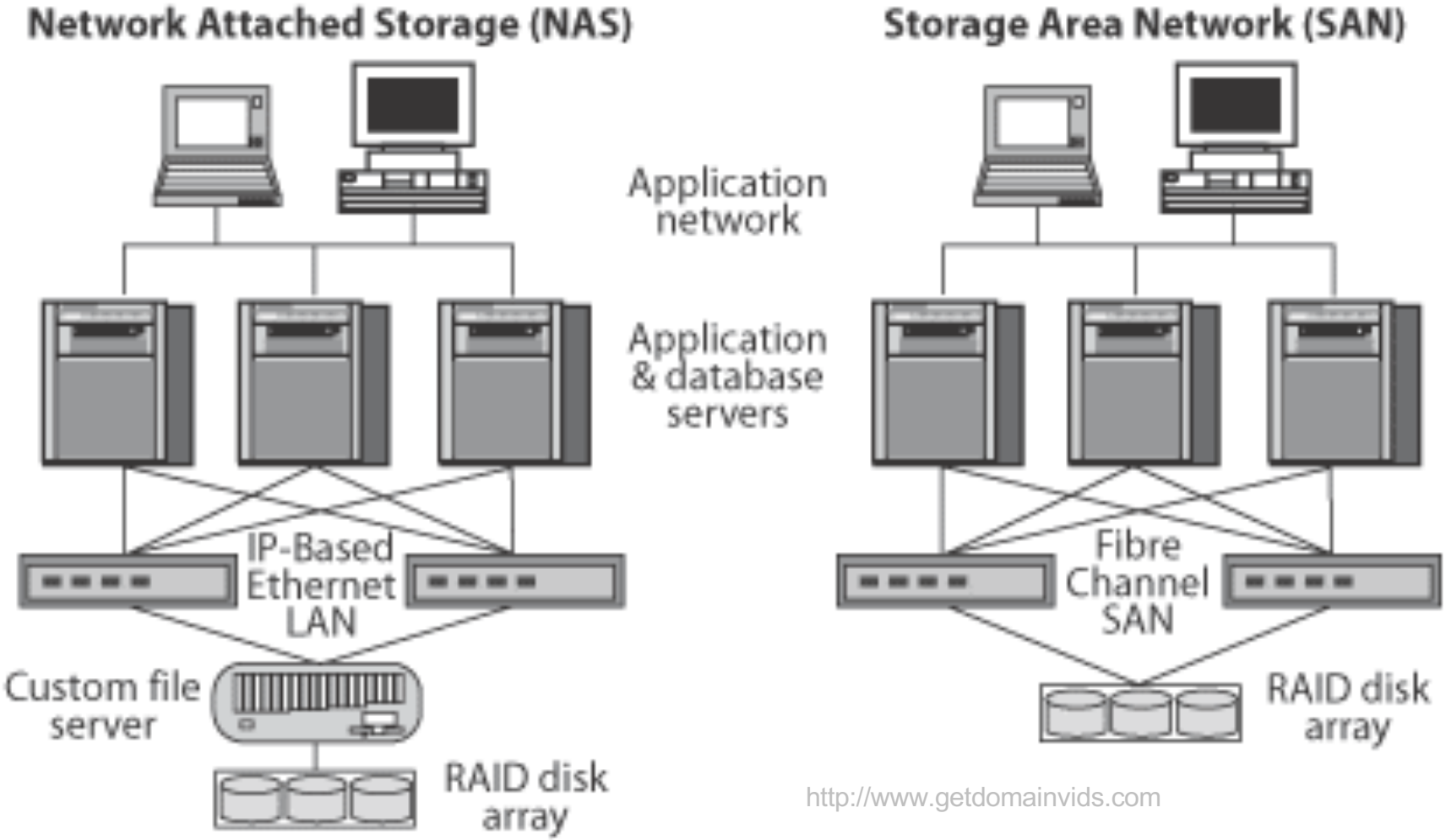
Evolução do storage distribuído/partilhado



https://sandipbagwe.wordpress.com

© Mário Antunes
mario.antunes@ipleiria.pt

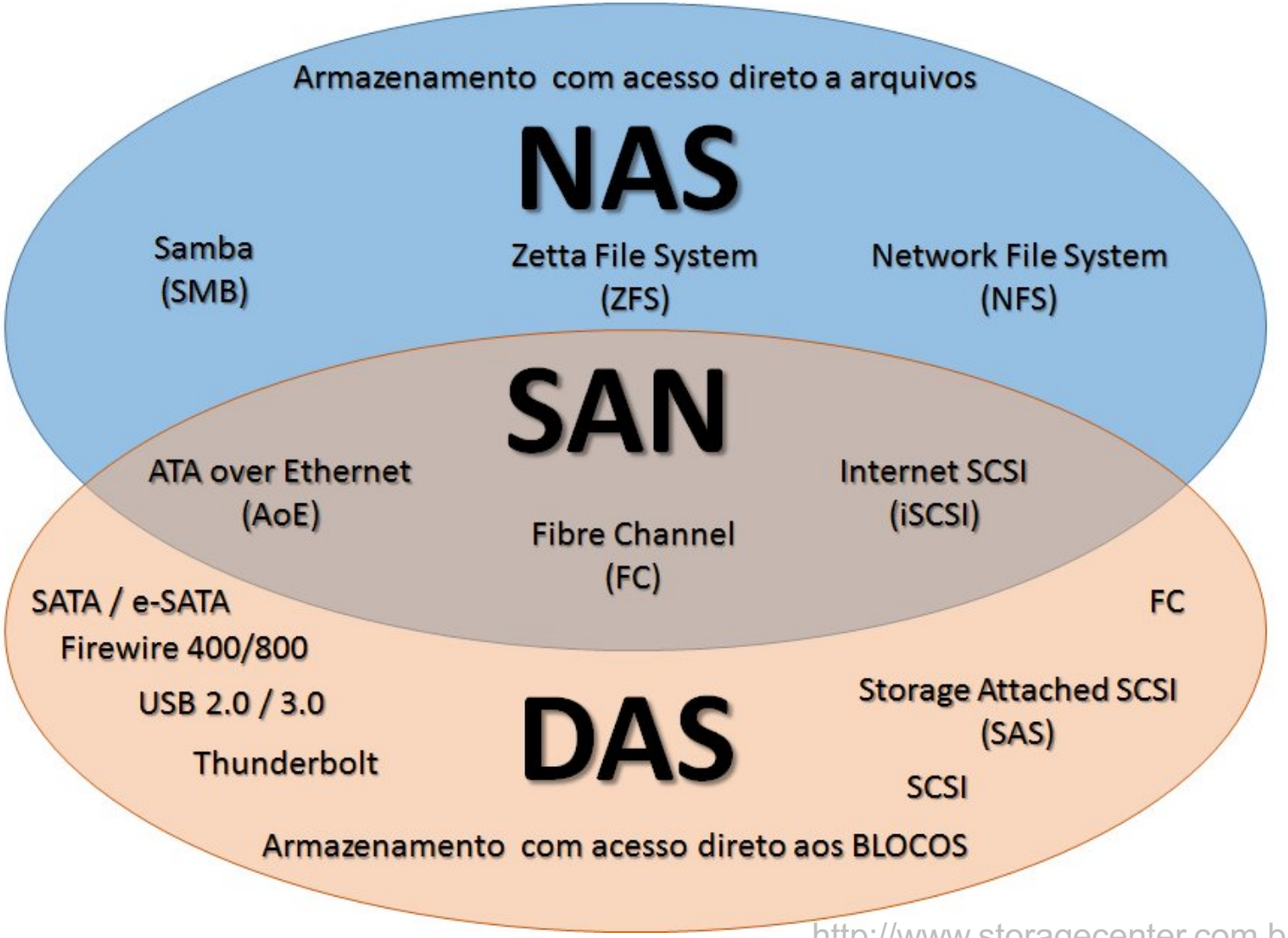
Evolução do storage distribuído/partilhado



http://www.getdomainvids.com

mario.antunes@ipleiria.pt

Evolução do storage distribuído/partilhado



http://www.storagecenter.com.br

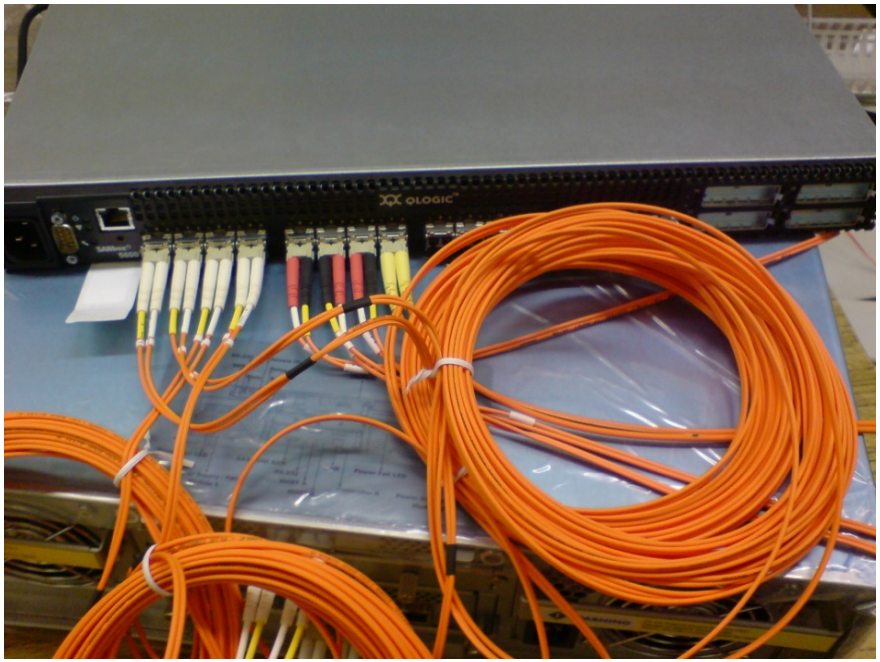
mario.antunes@ipleiria.pt

Hardware

Synology DS1813+ 8-Bay Scalable NAS

FC cable – Cat6

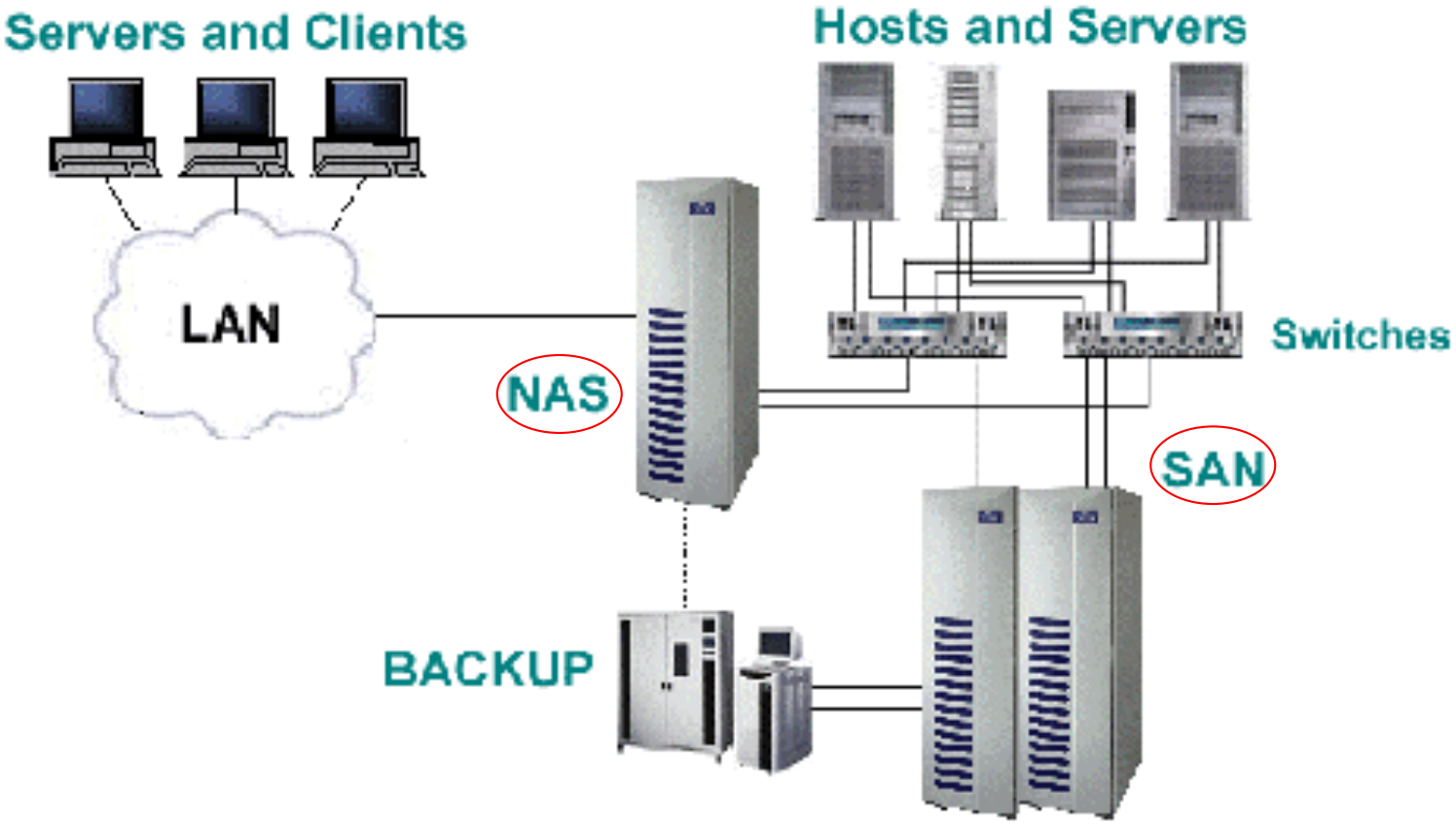
FC cable – optical fiber



Fiber Channel switch

mario.antunes@ipleiria.pt

Soluções híbridas - SAN / NAS



http://www.cityu.edu.hk

mario.antunes@ipleiria.pt

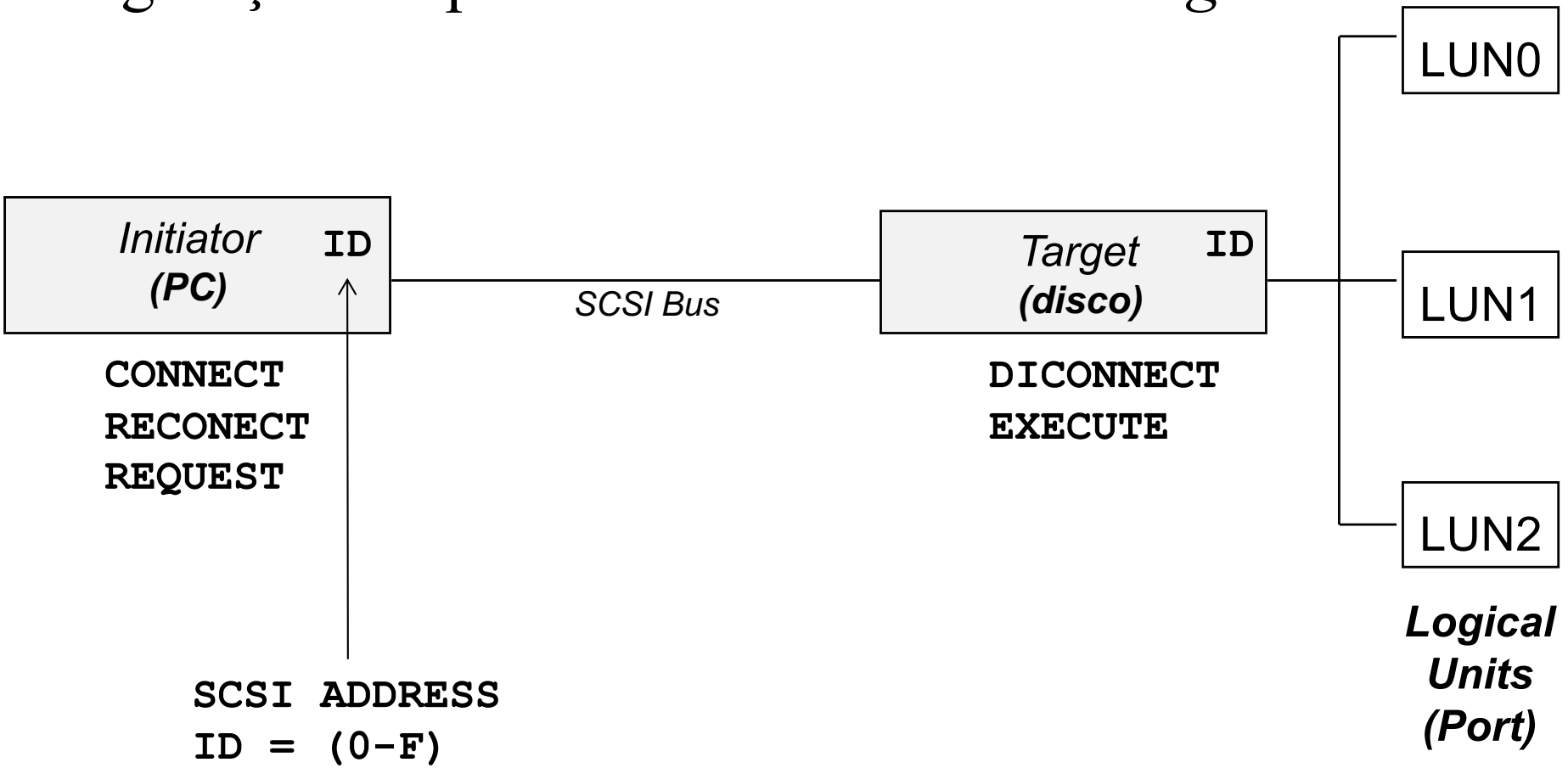
Tecnologia SCSI

- **S**mall **C**omputer **S**ystem **I**nterface
- Inicialmente desenvolvido pela Shugart Associates - SASI (Shugart Associates System Interface)
- Atualmente denominado SCSI e com um ANSI standard (T10)
- 3 versões: SCSI-1, SCSI-2 e SCSI-3
- Dispositivos comunicam através de um *bus*.
- Acesso dos hosts aos dispositivos: *block based*
- Principais limitações: comprimento (25 m); número de dispositivos suportados;

mario.antunes@ipleiria.pt

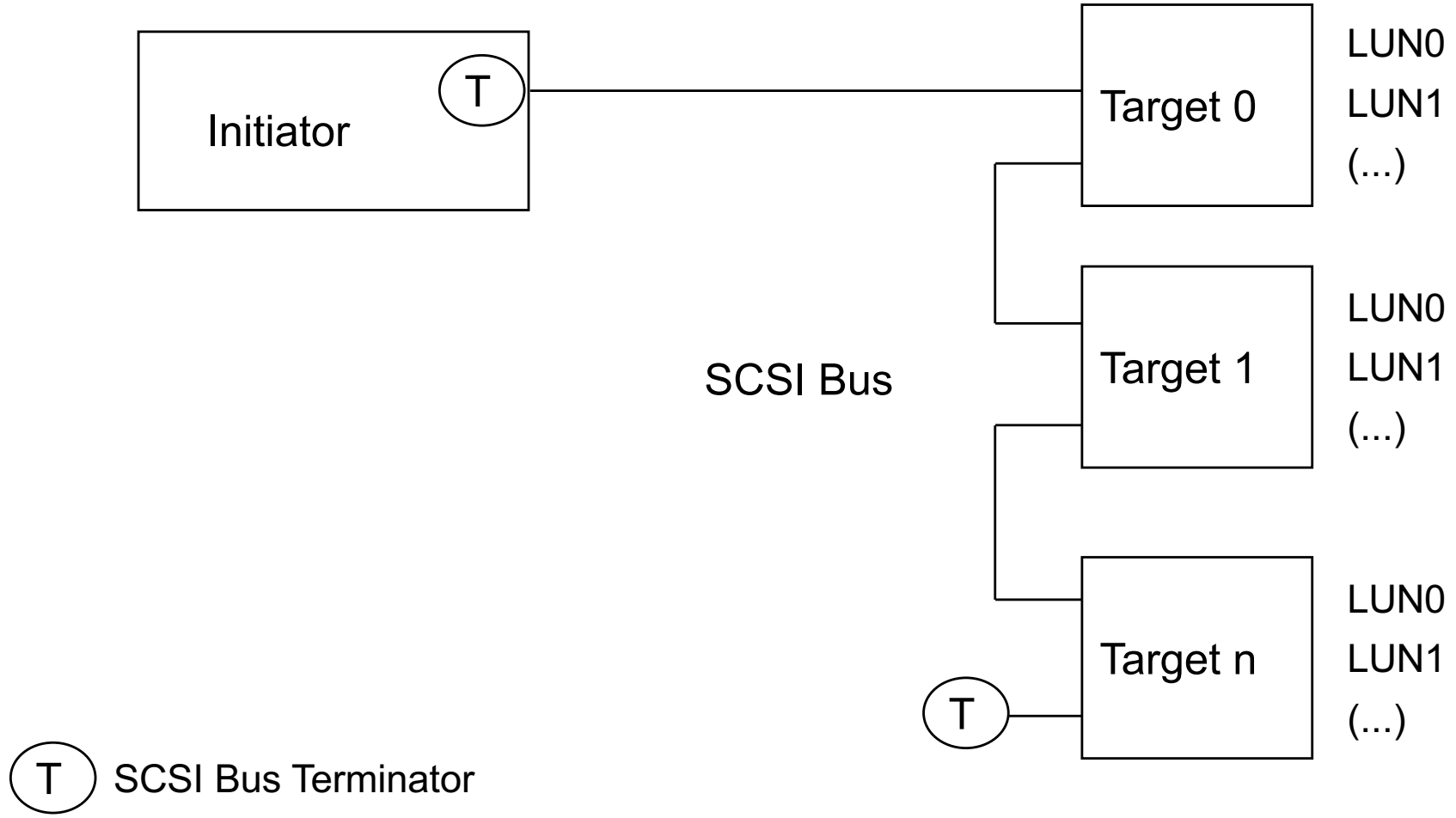
Tecnologia SCSI

Configuração simples: um *initiator* → um *target*



Tecnologia SCSI

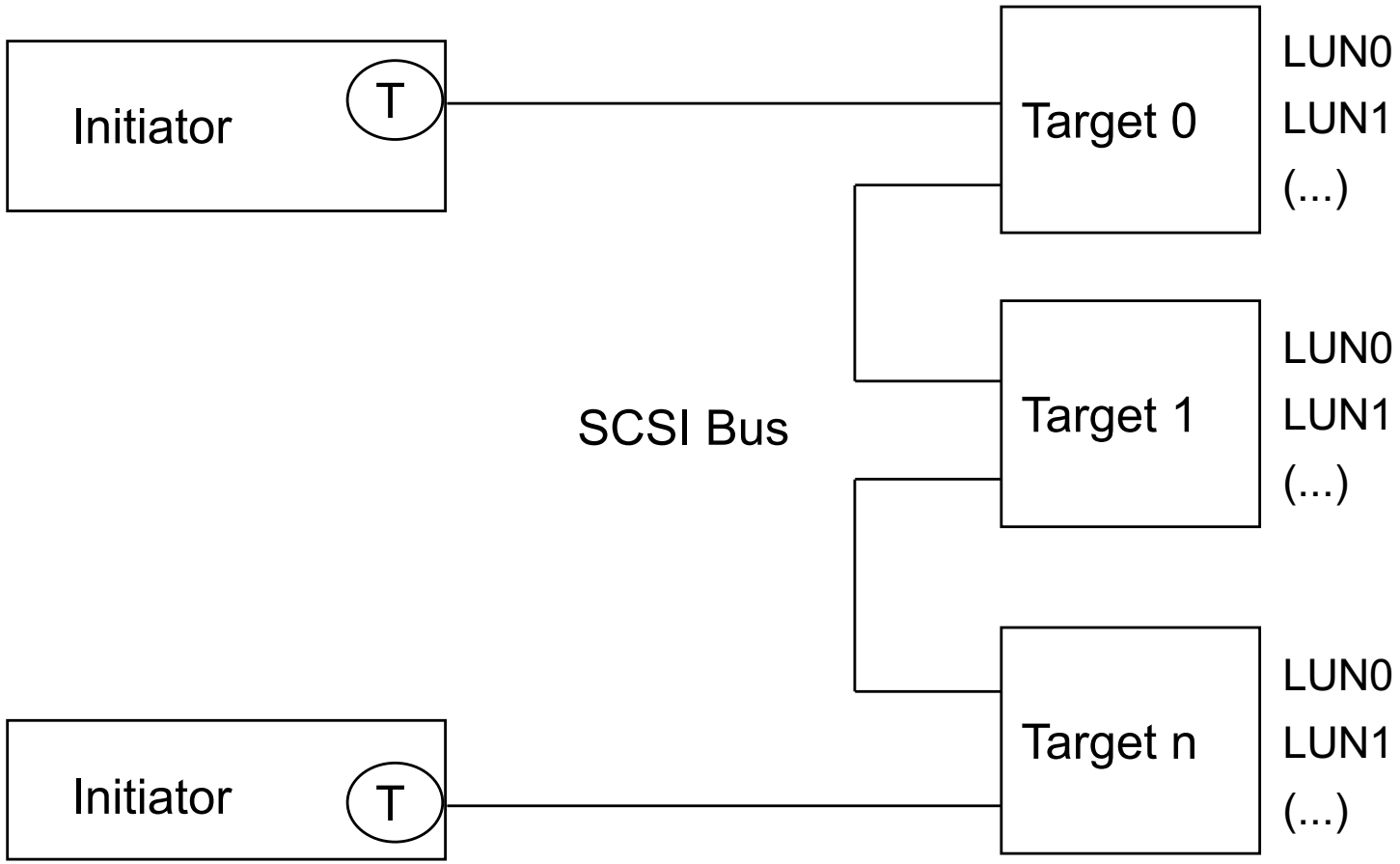
Configuração: um *initiator* → vários *target*



mario.antunes@ipleiria.pt

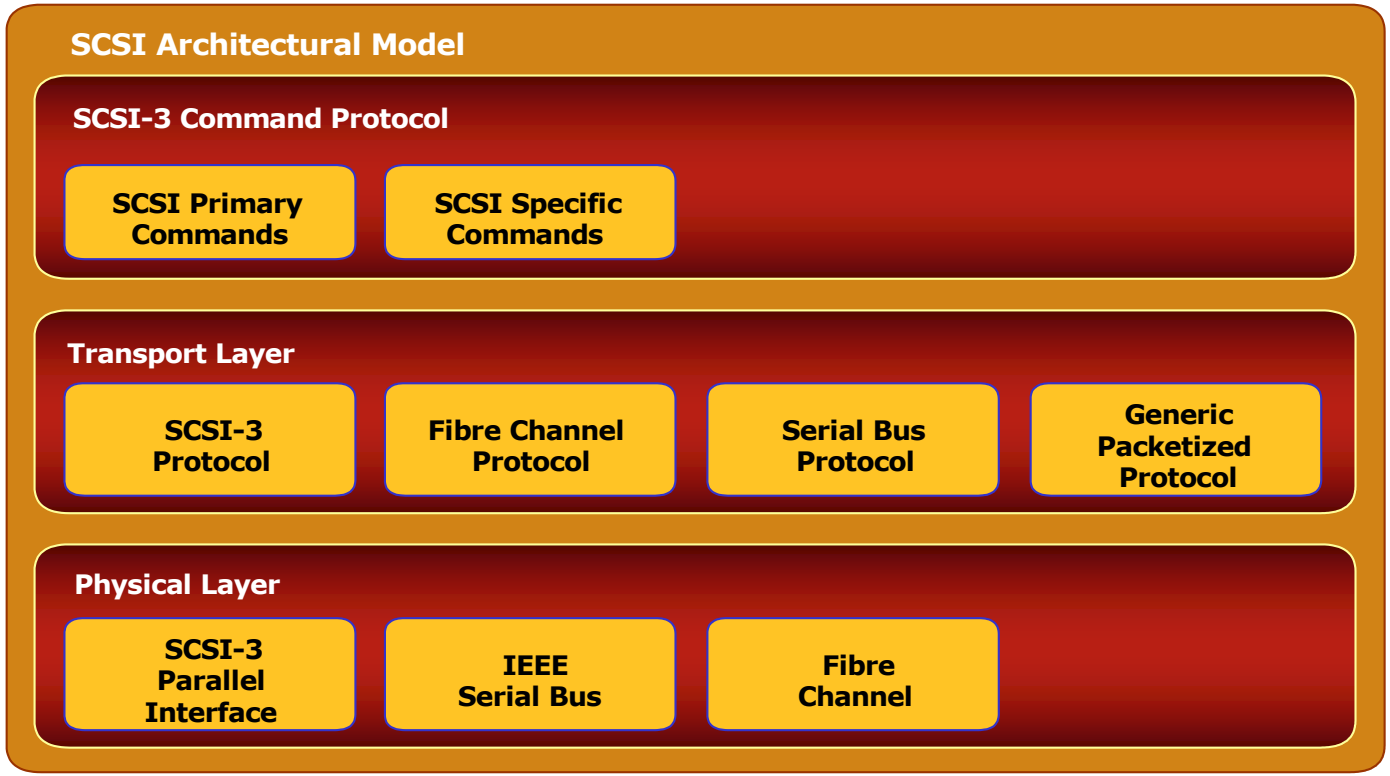
Tecnologia SCSI

Configuração: vários *initiators* → vários *target*



Tecnologia SCSI

Arquitetura



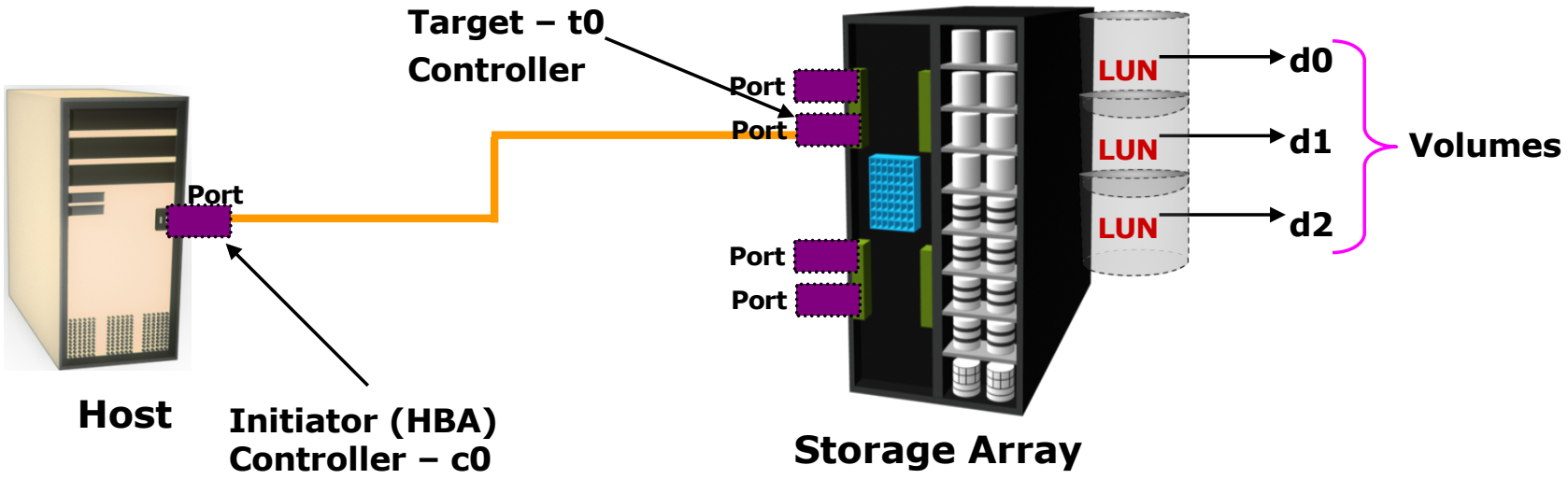
Comandos de I/O entre os dispositivos

Regras de comunicação entre os dispositivos

Detalhes da interface, adaptadores, etc..

mario.antunes@ipleiria.pt

Tecnologia SCSI - endereçamento



Host Addressing:

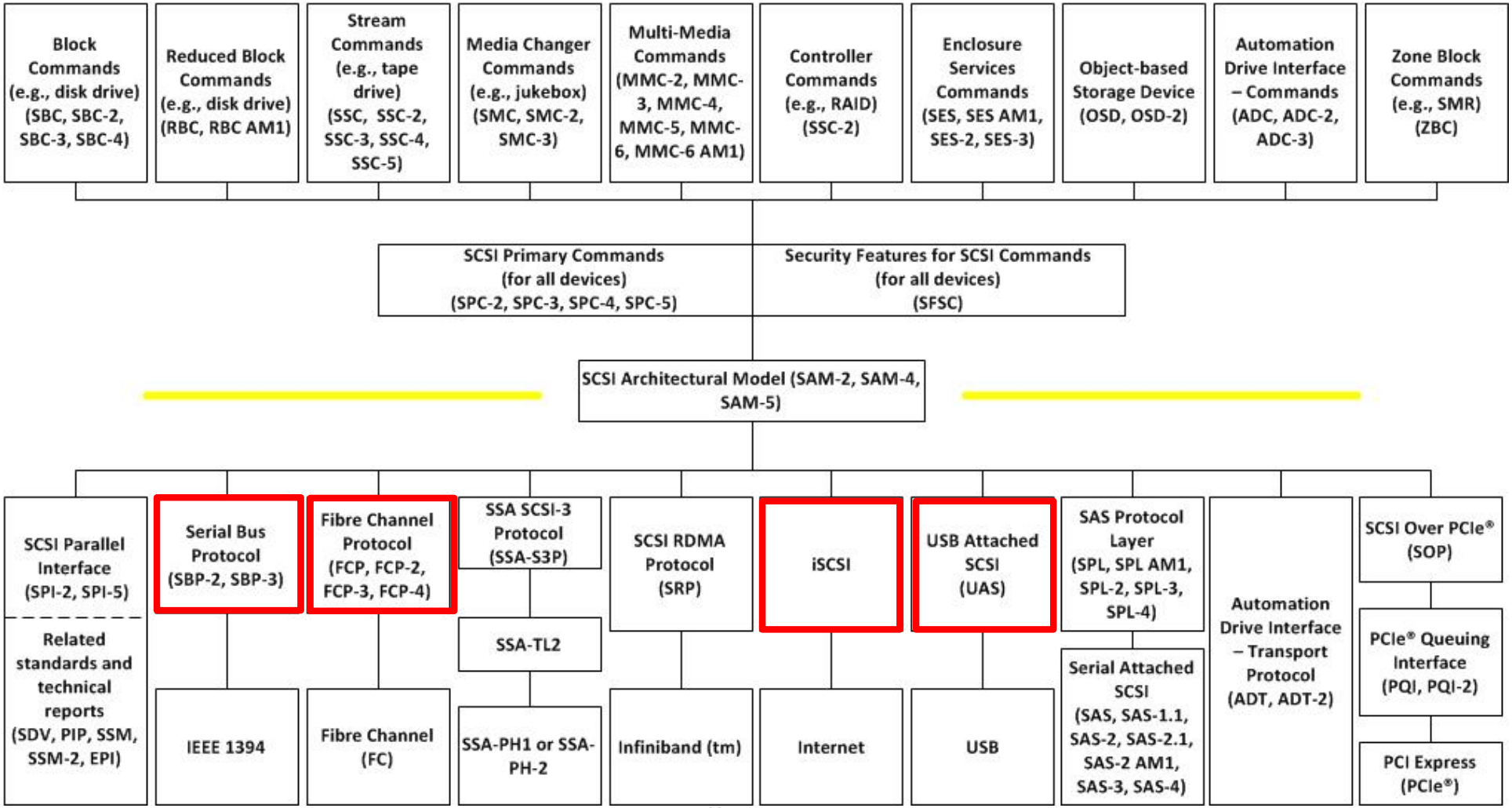
Storage Volume 1 - c0t0d0

Storage Volume 2 - c0t0d1

Storage Volume 3 - c0t0d2

Initiator ID	Target ID	LUN
c0	t0	d0

Tecnologia SCSI - standard



http://www.t11.org

mario.antunes@ipleiria.pt

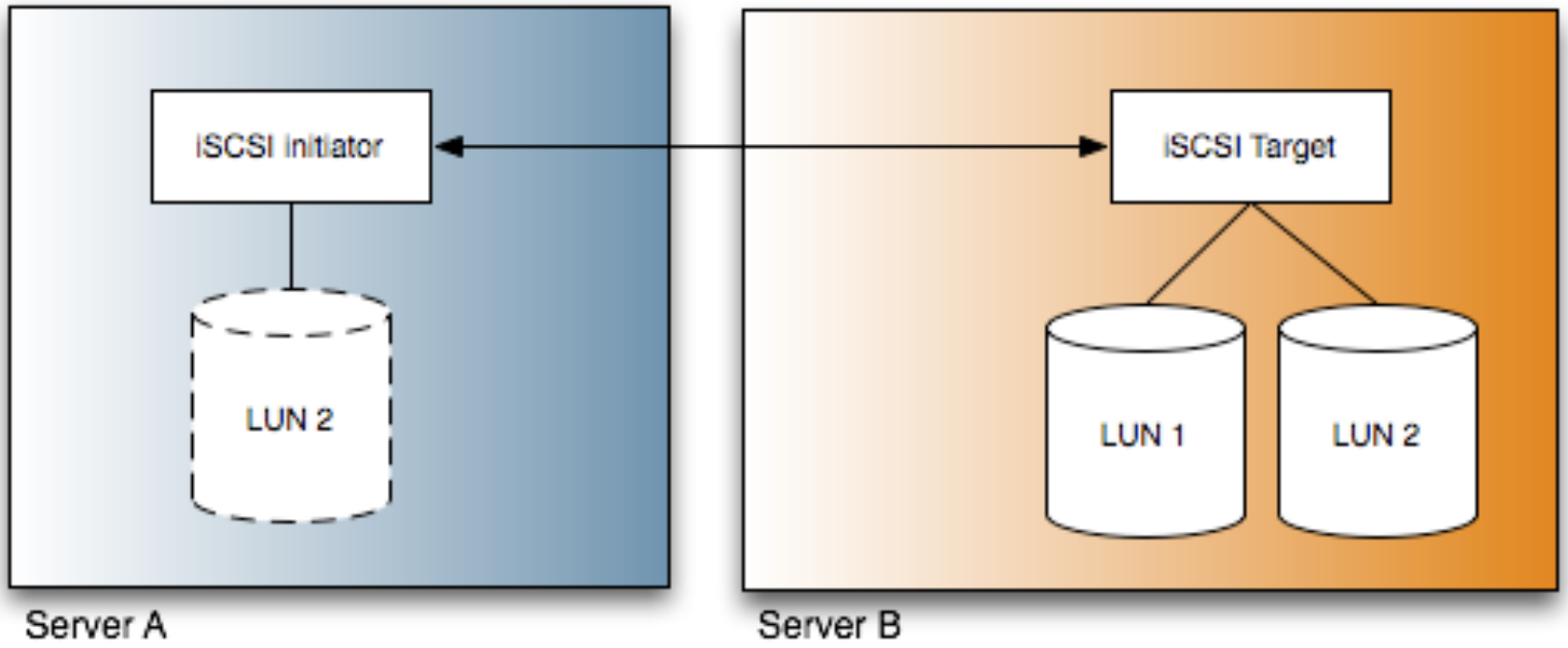
Tecnologia iSCSI

- **i**nternet **S**mall **C**omputer **S**ystem **I**nterface
- **Motivação:**
  - enviar comandos SCSI por rede TCP/IP (p.e. Internet).
  - acesso/partilha de storage através de longas distâncias
- Estabelece ligações *"initiator → target "* numa sessão TCP
- Topologia em estrela
- Normas IETF: RFCs 3720, 3721

mario.antunes@ipleiria.pt

Tecnologia iSCSI

Acesso remoto por iSCSI



Acesso a uma LUN remota *(target)* a partir de um dispositivo iSCSI *(initiator)*.

mario.antunes@ipleiria.pt

Tecnologia iSCSI

- **Connection** Ligação TCP usada para enviar mensagens de controlo, comandos SCSI, parâmetros e iSCSI PDUs. Poderá haver várias "connections" entre o *target* e o *initiator*, todas na mesma "session".

- **Session** Define um grupo de "connections" TCP a partir de um *initiator*. As "connections" podem ser adicionadas e removidas dinamicamente. O *initiator* consegue aceder a todas as "connections" numa sessão e o respetivo *target*.

```
1 session → n connections
```


mario.antunes@ipleiria.pt

Tecnologia iSCSI

Fases das sessões/ligações iSCSI:

- **LOGIN PHASE**
  - Estabelece ligação TCP
  - Autenticação de ambos os pontos da ligação
  - Negociação de parâmetros operacionais
  - Associação "connection → session"
- **FULL-FEATURE PHASE**
  - Transferência de dados

Tipos de sessões

- Normal
- Discovery

mario.antunes@ipleiria.pt

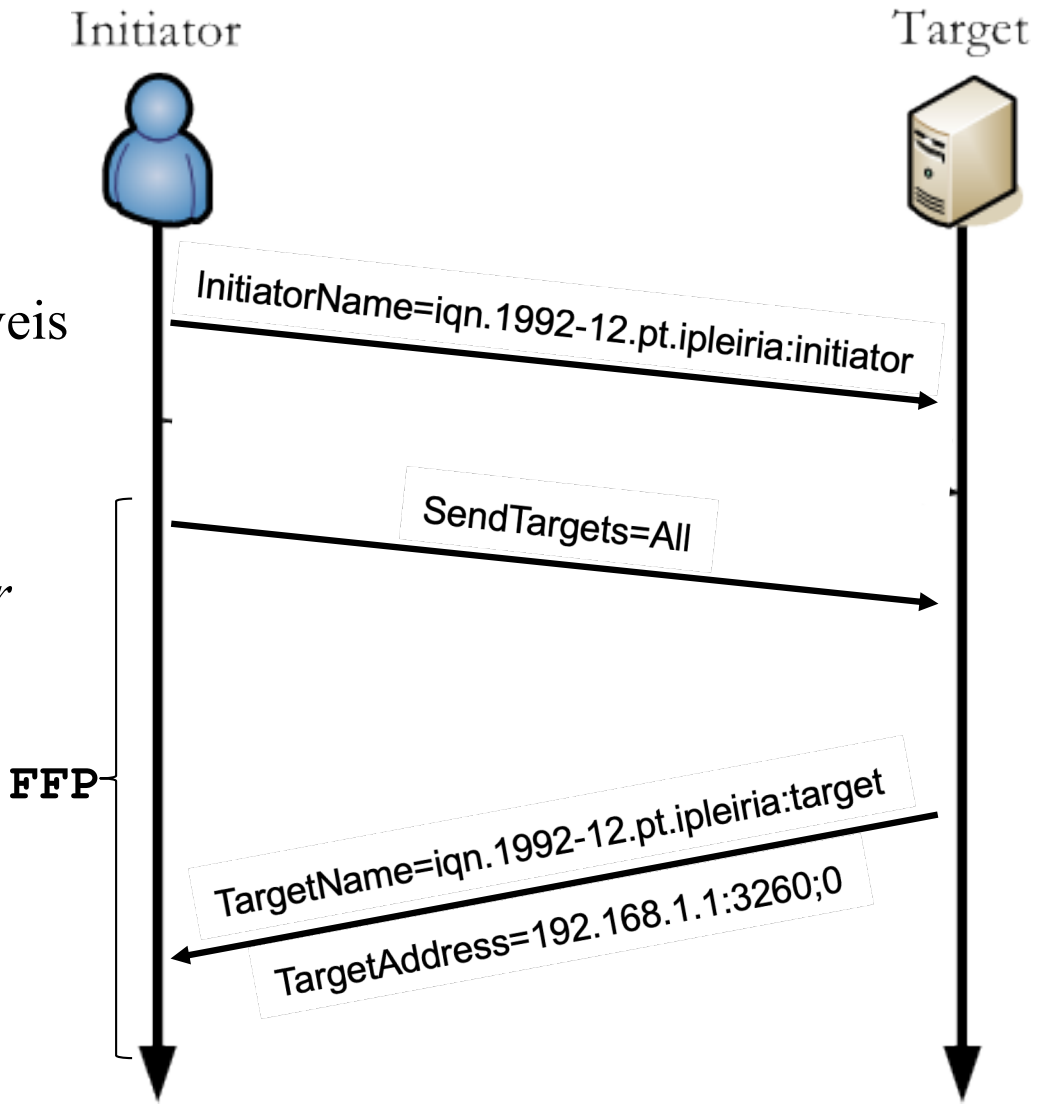
Tecnologia iSCSI

Sessão DISCOVERY

iSCSI *initiator* inicia procura de possíveis iSCSI *targets* a que se possa ligar.

Login Phase: Autenticação do *initiator* no *target* selecionado.

Full-Feature Phase: Troca de mensagens entre *initiator* e *target*.

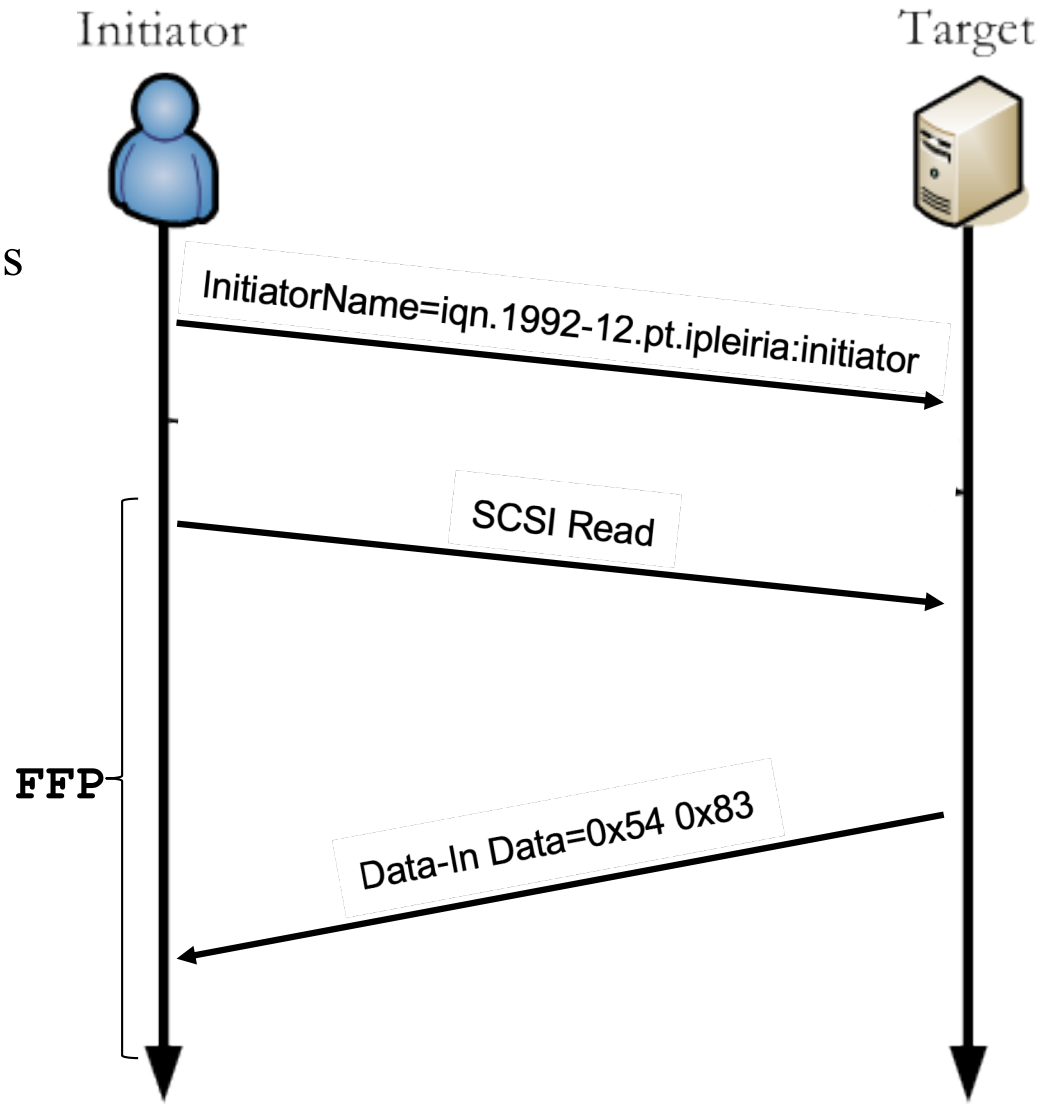


mario.antunes@ipleiria.pt

Tecnologia iSCSI

Sessão NORMAL

Initiator e *target* negoceiam parâmetros de comunicação (p.e. tamanho das mensagens individuais e número de sessões simultâneas.



Tecnologia iSCSI – convenção de nomes

RFC 3721 Apr 2004

Dois formatos principais: **`iqn`** e **`eui`**

iqn (iSCSI qualified Name): definir para cada dispositivo um nome único com detalhes sobre o dispositivo.

`Type  Date  Org.Unit  :  Location`

`iqn.2001-04.com.example:diskarrays-sn-a8675309`

Reverse DNS

Opcional

Tecnologia iSCSI – convenção de nomes

RFC 3721 Apr 2004

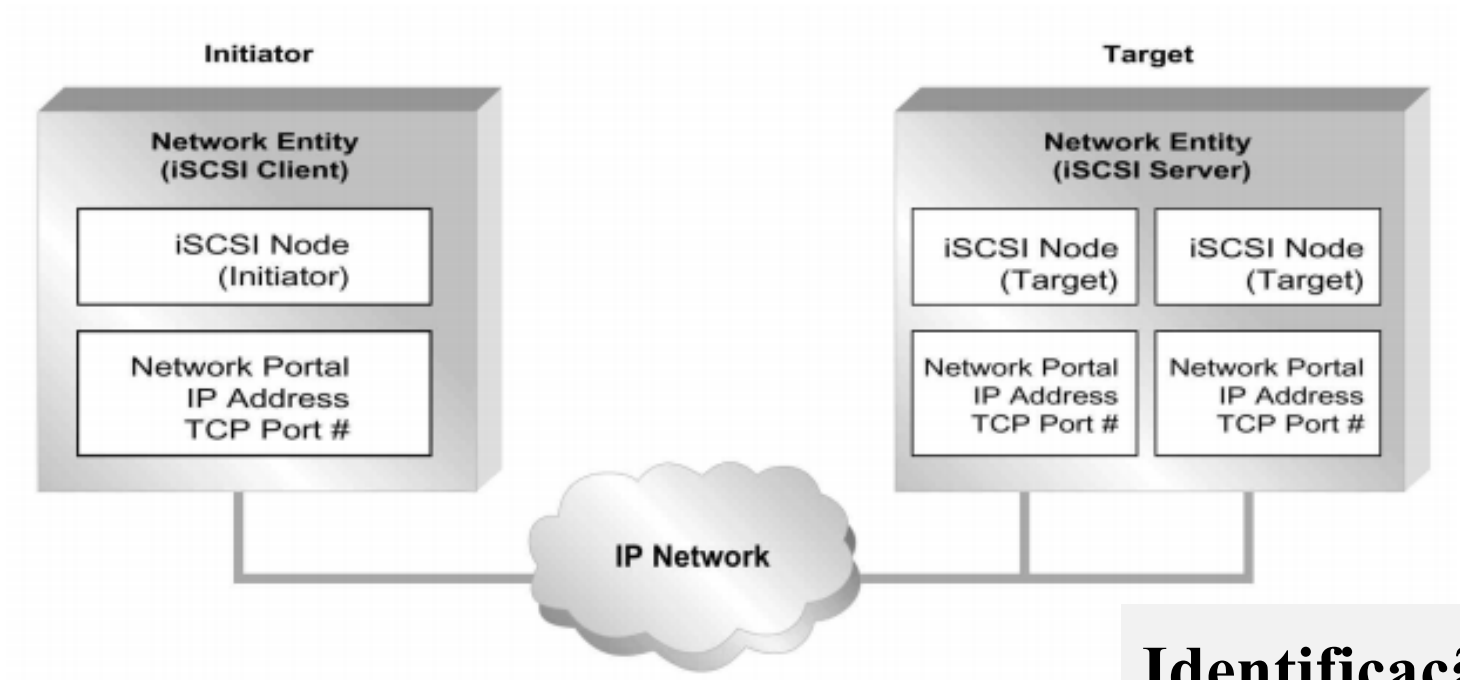
eui: utilizar a convenção IEEE EUI (Extended Unique Identifier) para a definição de IDs (hexadecimal)

Type	**EUI-64 identifier**
eui.02004567A425678D	

iSCSI alias: utilizar um nome apelativo para o dispositivo. Esta designação será utilizada durante a fase de login, entre o *initiator* e o *target*.

mario.antunes@ipleiria.pt

Tecnologia iSCSI – identificação dos pontos



iSCSI Technical White Paper; White paper; Nishan Systems

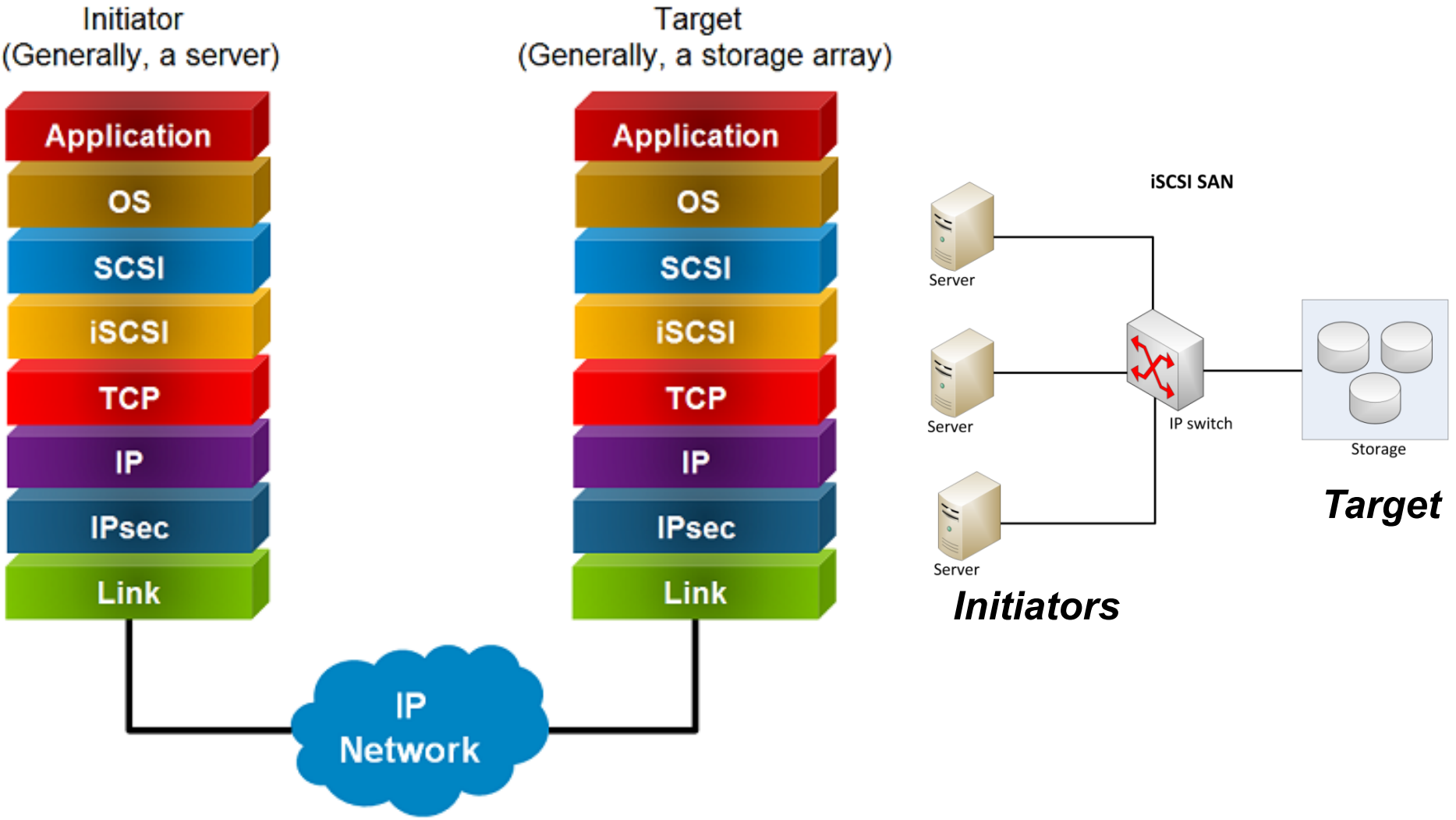
Normalmente:

TCP Port = 860 e 3260

Identificação completa

- hostname ou endereço IP
- TCP Port
- iSCSI ID
- CHAP password (opcional)

Tecnologia iSCSI

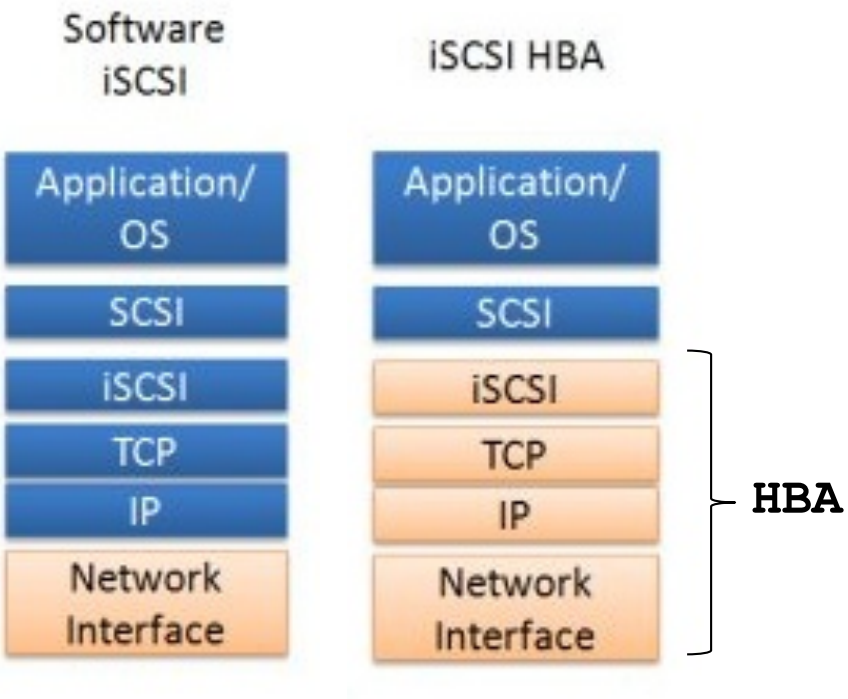


http://nextgencomputing.tumblr.com

mario.antunes@ipleiria.pt

Tecnologia iSCSI - implementações

Sistemas operativos disponibilizam software para target e/ou initiator.



Targets

Host Bus Adapter (HBA) instalados em discos (ou tapes).

Implementação para Linux:

- `http://linux-iscsi.sourceforge.net`
- `http://www.open-iscsi.org`

mario.antunes@ipleiria.pt

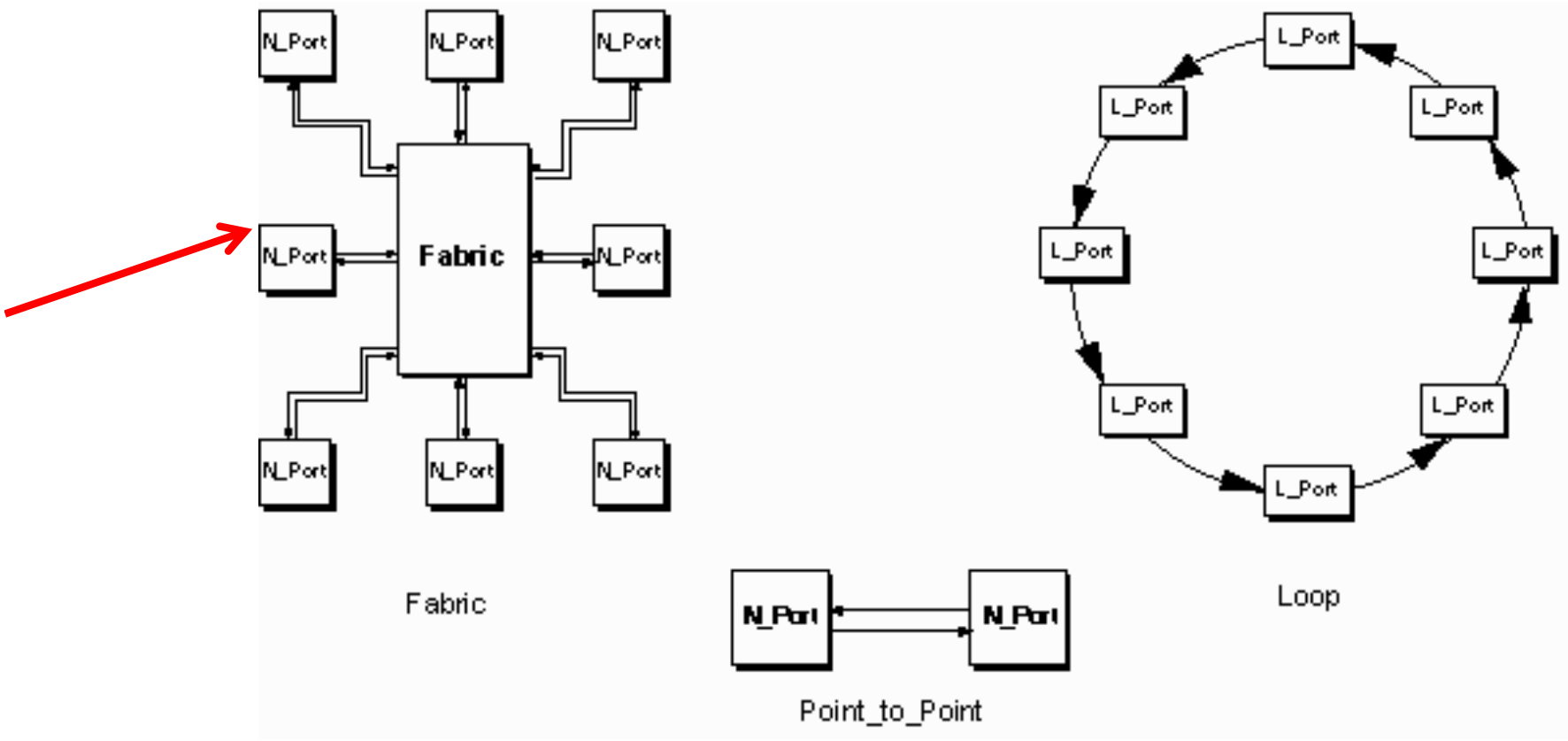
Tecnologia Fiber-Channel (FC)

- Tecnologia de redes para acesso massivo a storage
- Substituto natural do SCSI: mais rápido e maiores distâncias
- Utiliza o Fiber-Channel Protocol (FCP) , norma ANSI-T11
- Um dos interfaces mais usados para SAN (juntamente com iSCSI)
- Originalmente para fibra (~ 10Kms). Mais recentemente, cobre.
- Utiliza FC Protocol (FCP) na camada de transporte
- Interage com outros protocolos: p.e. IP e SCSI
- Noções similares ao SCSI: *initiator*, *target* e HBA.

mario.antunes@ipleiria.pt

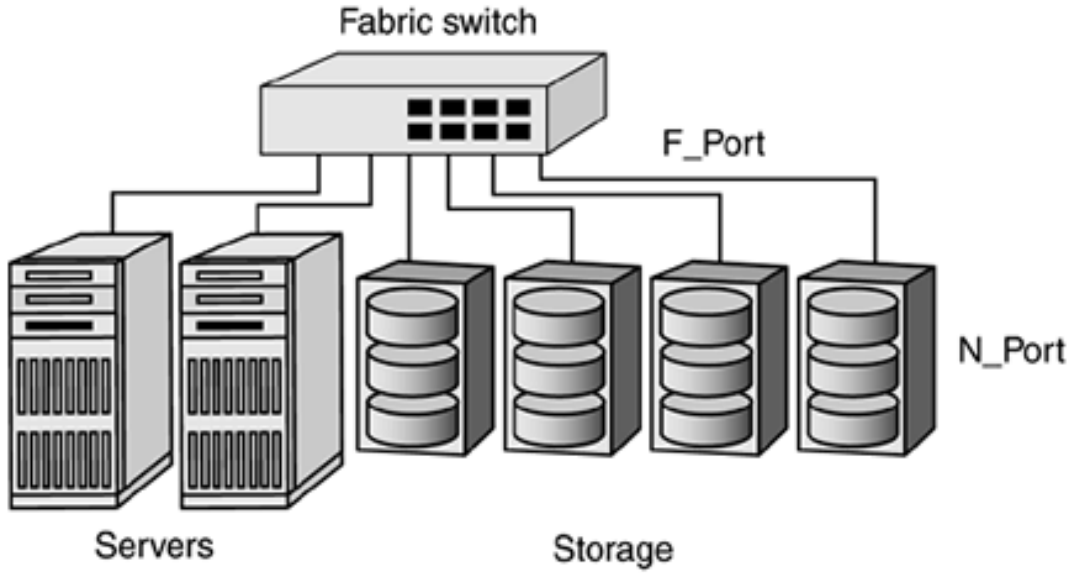
Tecnologia Fiber-Channel (FC)

Topologias disponíveis:



Tecnologia Fiber-Channel (FC)

Topologias do tipo *switch fabric*:



N_Port: porta de ligação de um nó FC ao switch

F_Port: porta do switch para ligar a um nó FC

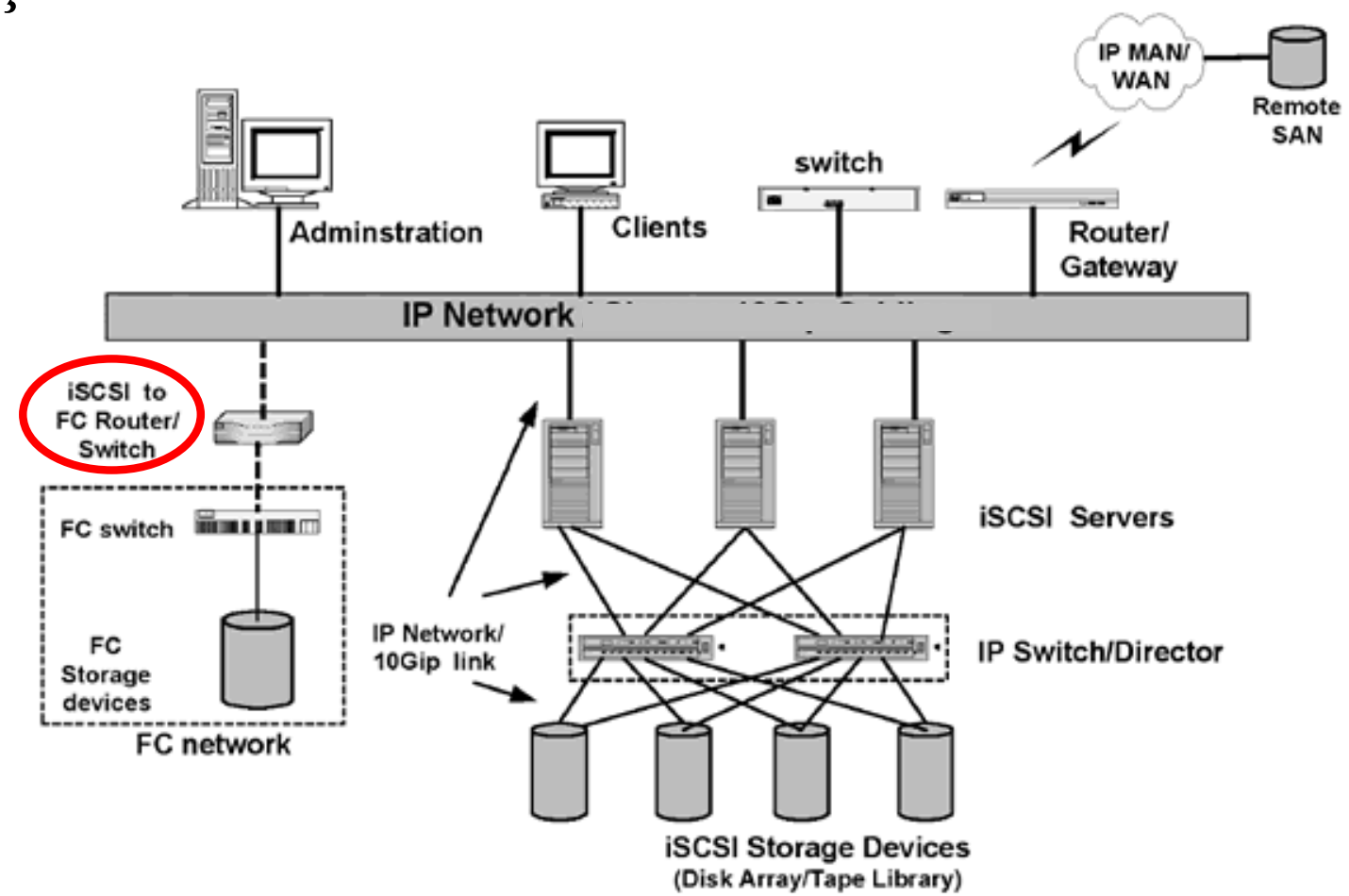
L_Port: porta usada para ligação de um nó a um FC loop

NL_Port: porta de um nó para ligação simultânea a um loop FC e a um switch

mario.antunes@ipleiria.pt

Tecnologia Fiber-Channel (FC)

Integração FC - iSCSI



www.siemon.com/

mario.antunes@ipleiria.pt

Outras tecnologias

- ATA-over-Ethernet (AoE)
- InfiniBand (IB)
- Fibre Channel over Ethernet (FCoE)
- Fibre Channel over IP (FCIP)
- HyperSCSI - SCSI over Ethernet
- iSCSI Extensions for RDMA (iSER)
- Internet Fibre Channel Protocol (iFCP)
- Serial Storage Architecture (SSA – IBM)

mario.antunes@ipleiria.pt

Conclusões

- Necessidades de replicação de dados é cada vez maior
- Um exemplo claro de clusters de HA associada à necessidade de balanceamento de carga
- Storage distribuído e virtualizado tomou maior interesse com a cloud
- A seguir de perto: cloud providers (Google et al.) e Apache (Apache Ecosystem)

mario.antunes@ipleiria.pt

Bibliografia

- Marcus E, Stern H., *"Blueprints for high availability";* 2003; *Wiley;* ISBN: 0471430269
- Luiz André Barroso, Jimmy Clidaras, Urs Holzle; *"The datacenter as a computer"*; Morgan and Claypool Editors; ISBN: 978-1627050098; 2013 [pdf]

mario.antunes@ipleiria.pt