Clusters de alta disponibilidade

UFRGS
Taisy Silva Weber

Cluster

- cluster ou agregado
  - computadores com múltiplos processadores
    - termo usado para vasta gama de configurações
      - número variável de nodos de computação convencionais: de 2 nodos a poucos milhares
      - opcionalmente alguns dispositivos de armazenamento compartilhados
      - interconexões de alta velocidade

exemplo de arquitetura tolerante a falhas
exemplo da aplicação de conceitos de sistemas distribuídos

The Design and Architecture of the Microsoft Cluster Service - A Practical Approach to High-Availability and Scalability, Werner Vogels et al., IEEE 1998

Definição

- **coleção de computadores** que trabalham visando prover um sistema de grande capacidade.
  - deve ser tão **fácil de programar** e de gerenciar como um único computador de grande porte.
- **vantagens**
  - pode **crescer muito** mais do que um único computador (escalabilidade)
  - pode **tolerar defeitos em nodos** e continuar a oferecer serviços (*failover*)
  - pode ser construído a partir de **componentes de baixo custo**

Sistemas distribuídos versus cluster

- cluster são sistemas distribuídos
  - sem memória compartilhada
  - sem relógio global
  - comunicação por troca de mensagens
  - mas tem a vantagem da proximidade física
- técnicas de TF em sistemas distribuídos são úteis em clusters
  - comunicação de grupo e membership
  - checkpointing, logging e recuperação
  - tratamento de particionamento

Tipos

- implementação
  - por hardware: mais eficiente, pouco adaptável
  - por software: menor custo
- objetivos
  - alto desempenho
  - balanceamento de carga
  - alta disponibilidade

alguns autores falam de mais um tipo: **disponibilidade contínua**

vários objetivos podem ser **combinados**

Combinações de tipos

- bons esquemas de balanceamento de carga podem contribuir para aumentar a disponibilidade
- em cluster de alto desempenho:
  - nodos críticos podem compor um núcleo de alta disponibilidade
  - todos os nodos podem contribuir mantendo réplicas de dados ou processos, checkpoints e logs uns dos outros
- redundância inerente no cluster facilita implementar tolerância a falhas

HA-Cluster

- alta disponibilidade
  - tempo de inicialização após falha (*failover*) pode variar de poucos minutos até uma hora
  - aplicações em sistemas de missão crítica
  - servidores primário e backups
- disponibilidade contínua
  - tempo de *failover* na ordem de 10 segundos

primário e backup executam mesmos processos (warm backup)

Compartilhamento de disco

- sistemas de disco compartilhado:
  - necessitam de um gerenciador de bloqueio
    - evitar conflitos devido a requisições de acesso simultâneo a arquivos
      - um arquivo sendo escrito por um nodo não pode ser aberto para escrita em outro nodo
- sistemas de armazenamento não compartilhado:
  - cada nodo é independente
  - toda a interação é por troca de mensagens

Sinal de vida (*heartbeat*)

- mensagem periódica enviada de um processo a outro para indicar que continua operacional
  - detecção de falhas: ausência de heartbeats

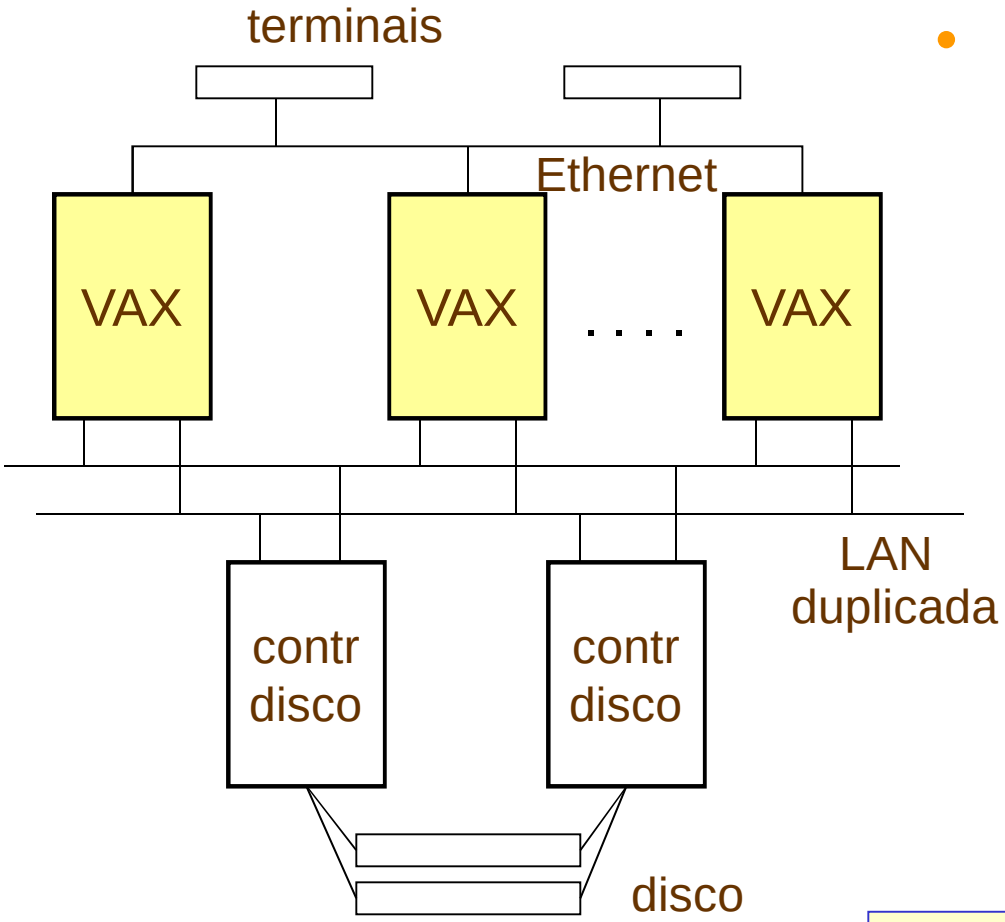
    heartbeats são esperados a cada poucos segundos

  - modelo fail-stop

    assume que se um nodo pára de enviar sinais, ele efetivamente não envia mensagens, nem altera dados no armazenamento estável

  - técnica antiga
    - muito usada antes mesmo dos primeiros clusters (Tandem,...)

Arquitetura VAX Cluster



- VAXcluster da Digital
  - primeiro cluster de sucesso
  - formado por nodos VAX
- se um VAX colapsa
  - todos os processos nele caem
  - serviços precisam ser reiniciados em outro servidor do cluster

não é transparente ao usuário

tempo longo de recuperação

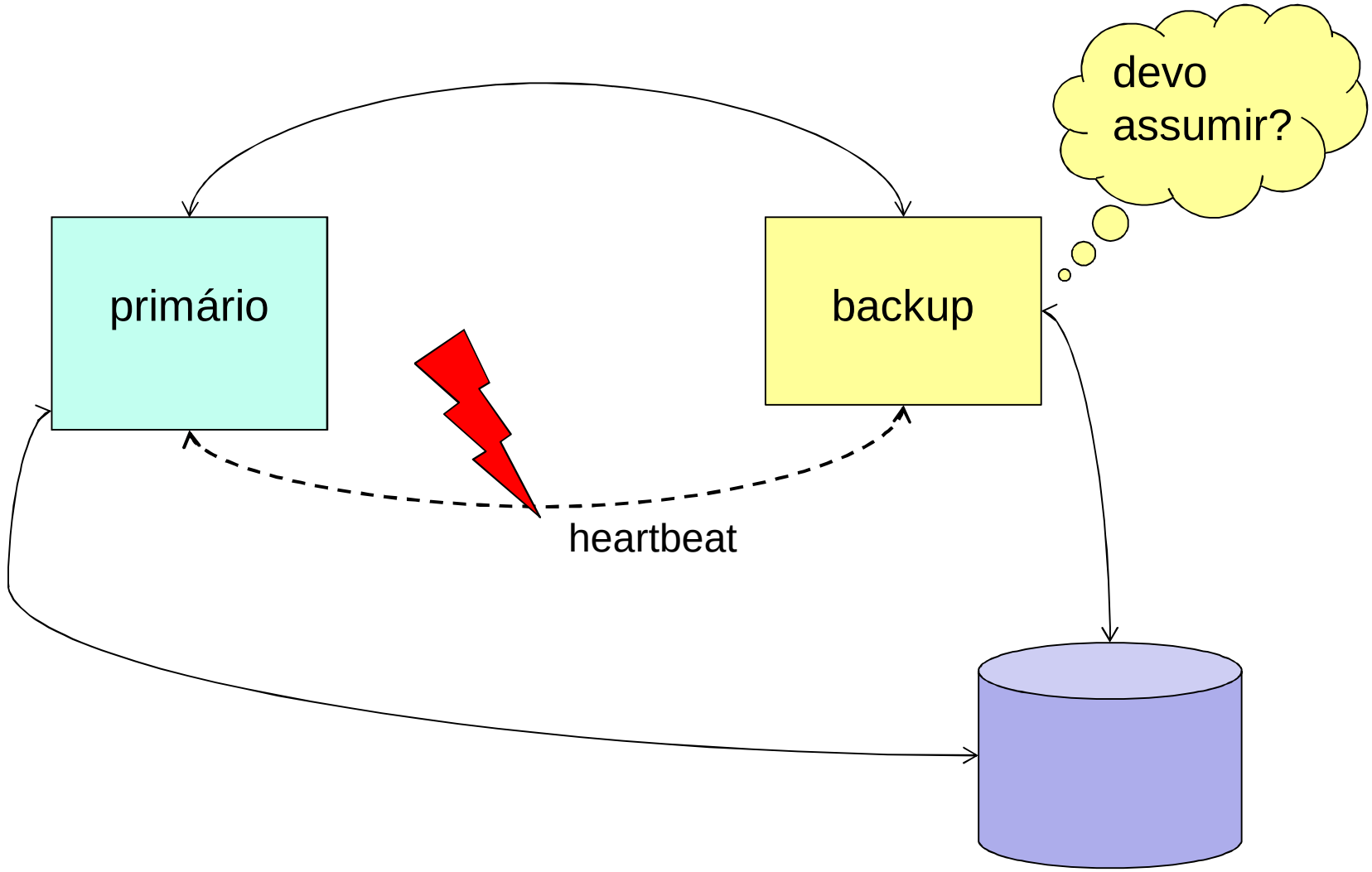
Disponibilidade em HA-clusters

- qual a disponibilidade efetivamente alcançada?
  - promessa de 99,99%
    - o VAXCluster não chegava a isso
- como avaliar?
    - experimentalmente por injeção de falhas
    - analiticamente através de modelos
    - ou durante operação levantando registros de falha (em logs por exemplo) e analisando

Problemas

- split-brain
  - um computador detecta o outro como defeituoso e assume as funções de primário
  - modelo fail-stop
    - assumido pelos fabricantes mas raramente implementado
  - particionamento

split brain



Bibliografia

- Birman, K. *Building secure and reliable network applications.* Manning Publications Co, Geenwich, 1996
- Vogels, W. *The Design and Architecture of the Microsoft Cluster Service - A Practical Approach to High-Availability and Scalability*, FTCS-IEEE, 1998
- Azagury, Alain et al. *Highly Available Cluster: a Case Study*. FTCS-IEEE, 1994
- Hughes-Fenchel, Gary. *A Flexible Clustered Approach to High Availablity*. FTCS-IEEE, 1997
- links de fabricantes