Relatório de Incidente
Interrupção no funcionamento do sistema de armazenamento de dados da FEEC

Campinas, 30 de maio de 2016.

1. Resumo:

Este relatório descreve os fatos relacionados à falha ocorrida no dia 13/04/2016 no sistema de armazenamento de dados (*storage*) utilizado pela FEEC. Lembrando que estes são os fatos conforme puderam ser observados pela equipe da DTI/FEEC, sendo que o encaminhamento de um relatório técnico por parte da equipe de suporte do fabricante ainda está pendente.

2. Descrição geral do equipamento:

O sistema de armazenamento de dados é composto por um equipamento da marca Huawei, modelo OceanStor S2600. O S2600 é um *storage* com suporte a discos SAS/SATA II, com controladoras e fontes de alimentação redundantes. O módulo (bandeja) principal inclui duas controladoras e duas fontes, bem como acomoda 12 (doze) discos. O sistema pode ser expandido através de bandejas extras para discos (Huawei modelo D120S), interligadas às controladoras através de enlaces SAS, configurados em *daisy-chain* para efeito de redundância. Cada bandeja D120S conta com duas fontes de alimentação redundantes e comporta também 12 discos.

Cada controladora provê 4 (quatro) portas GigabitEthernet para conectividade com as máquinas clientes (servidores de aplicação) e 2 (duas) portas SAS para conexão entre as bandejas de expansão.

A configuração existente atualmente na FEEC consiste de uma unidade S2600 associada a 3 (três) bandejas D120S, utilizando discos SATA II de 1TB ou 2TB, somando um espaço bruto de armazenamento de 75TB.

A unidade controladora S2600, juntamente com uma bandeja D120S, foi instalada em Abril/2011, sendo uma segunda bandeja incluida em Outubro/2012 e a última em Fevereiro/2015.

Os discos são associados internamente em grupos RAID nivel 5 (suporte à perda de um disco por grupo sem interrupção do acesso aos dados). Cada grupo RAID é subdividido em unidades lógicas (*LUNs*) que por sua vez são disponibilizadas para as diversas máquinas clientes através de conexões iSCSI.

O espaço de armazenamento é utilizado tanto para dados de usuário (e-mail, servidor de arquivos, etc.), como para prover os discos para as máquinas virtuais que rodam os diversos servidores da rede da FEEC. Alem disto o backup destes dados é realizado também neste storage.

3. Descrição do incidente:

A partir do final de Março/2016 começaram a se manifestar eventos aleatórios de perda de performance no storage, sendo o sintoma mais evidente a perda de conexão por alguns minutos entre os clientes iSCSI e as controladoras. Estes eventos ocorriam em intervalos variados, as vezes duas ou três vezes ao dia, as vezes espaçados de vários dias.

Em nenhum destes eventos foi gerado qualquer registro de erro pelo software de gerenciamento do storage. Como relatos similares envolvendo storages de fabricantes diferentes podem ser encontrados em diversos fóruns técnicos, a interpretação inicial foi a de que se tratava de limitações de desempenho do ambiente de rede e foram adotadas medidas para contornar estas limitações, tais como remanejar as conexões de rede usadas pelos clientes iSCSI para aumentar a banda disponivel, etc.

Finalmente, por volta das 15h00 do dia 13/04/2016, quarta-feira, ocorreu novamente perda de conexão com o storage, mas desta vez o sistema de gerenciamento apontou um erro crítico, identificado nas mensagens de erro como "*coffer disk failure*".

Uma parte de cada um dos quatro primeiros discos da bandeja que contém as controladoras é utilizada para armazenar o sistema operacional do storage (essencialmente uma versão de sistema operacional Linux modificada de forma proprietária pelo fabricante). Estes quatro discos são denominados *coffer disks*. De acordo com a documentação da Huawei, os *coffer disks* são agrupados em dois pares, sendo

que cada par consiste em um volume RAID-0, e que os pares são espelhados entre si usando RAID-1. Desta forma, em teoria, o sistema suportaria a perda de um par sem prejuizo para a operação.

No caso deste incidente não foi o que se verificou: a perda de um único *coffer disk* resultou na interrupção dos serviços. A detecção de um erro critico fez com que o sistema operacional colocasse todos os 48 discos do sistema em modo *offline*, supostamente para prevenir erros mais graves. Consequentemente os volumes RAID ficaram indisponíveis e as máquinas clientes perderam acesso aos discos via iSCSI.

Como os servidores virtualizados se utilizam destes discos, esta perda de conexão, além de tornar os dados de usuário inacessíveis, resultou na interrupção da execução destes servidores. Sendo que atualmente 100% dos servidores de rede da FEEC são virtualizados, o impacto foi significativo.

O sistema de gerenciamento da Huawei não disponibiliza acesso de "baixo nivel" para correção de problemas como o descrito acima, de modo que é necessário abrir um chamado no suporte para que uma equipe técnica do fabricante faça o diagnóstico e reparo, em geral remotamente.

Como a FEEC não conta com contrato válido de suporte para este equipamento, a DTI contatou (ainda na tarde do dia 13/04) a empresa InterQuattri Informática e Telecomunicações Ltda., que foi o fornecedor do storage. Através da InterQuattri foi possivel a abertura de um chamado de suporte em carater emergencial, condicionado à efetivação futura de um contrato de suporte.

Em função desta exigência e das tratativas necessárias entre a Huawei, InterQuattri e a Diretoria da FEEC, o atendimento ao incidente se iniciou efetivamente por volta das 10h00 do dia 15/04 (sexta-feira), através de uma equipe técnica da Huawei.

Na ausência de um relatório por parte da Huawei sobre o que foi realizado, podemos apenas relatar o que pôde ser observado pela equipe da DTI: interagindo remotamente com o sistema de gerenciamento do storage, em modo de "depuração" (o qual é protegido por senha de conhecimento somente do fabricante), os técnicos da Huawei recolocaram cada disco novamente *online* e em seguida ativaram os volumes RAID correspondentes.

Este procedimento se estendeu até a madrugada do sábado (16/04) e incluiu em mais de um momento a coleta de diversos logs do ambiente (aos quais a DTI ainda não teve acesso), bem como, ao final, a atualização do *firmware* do equipamento.

Com exceção da uma única interveção física para remover e em seguida reinserir um dos discos na bandeja, não foi feita nenhuma alteração no hardware do storage. O disco de *coffer* que apresentou problema pôde ser reparado sem necessidade de substituição.

O sistema foi liberado pela equipe da Huawei na manhã do dia 16/04 para que o pessoal da DTI pudesse "recolocar as aplicações no ar, para efeito de teste". Esta etapa foi completada por volta das 17h00 deste mesmo dia.

Do ponto de vista do storage o incidente não resultou em perda de informações, ou seja, toda a configuração do ambiente, incluido a definição dos volumes RAID, particionamento destes volumes em LUNs, permissões de acesso às LUNs pelos clientes iSCSI, etc., foi preservada.

No entanto do ponto de vista das máquinas clientes os sistemas de arquivo definidos sobre estas LUNs sofreram inconsistências, com graus variados de severidade em função do número de acessos pendentes para cada LUN no momento em que o sistema ficou *offline* no dia 13/04.

Todos os dados armazenados naquele momento nos caches das aplicações e/ou no cache das controladoras do S2600 não foram salvos em disco. Consequentemente ao reiniciar os servidores e realizar as checagens de integridade dos sistemas de arquivo, foram observados problemas como perdas em arquivos abertos naquele momento e conflitos de propriedade e/ou permissões. Tais problemas foram mais perceptíveis no sistema de e-mail, justamente por ter um ciclo de escrita/deleção mais frequente do que outras aplicações.

A maioria destes problemas foi corrigida nos dias subsequentes, eventualmente com a restauração de arquivos do backup.

4. Considerações finais:

Conforme destacado anteriormente, estas são as informações disponíveis até o momento. A identificação das causas exatas da falha de storage depende de um relatório mais detalhado, o qual provavelmente será disponibilizado pela Huawei somente após firmado o contrato de suporte, o qual ainda está em tramitaçao.

Da mesma forma, na ausência deste detalhamento, não podemos inferir se as medidas adotadas pela equipe de suporte foram corretivas, no sentido de prevenir a reincidência da falha, ou meramente paliativas.

De qualquer maneira, podemos apontar alguns questões relacionadas à disponibilidade ou robustez deste ambiente que precisariam ser consideradas pela FEEC:

- O sistema S2600, a despeito de ainda estar funcional (considerando que as causas que levaram a esta falha foram corrigidas), já conta com limitações de desempenho, na medida em que não oferece suporte a discos mais rápidos (tais como SATA III, 6Gbps ou NL-SAS, 12 Gbps) nem conexões de rede 10 GbE.
- Este produto já foi descontinuado pela Huawei e, segundo informações do fornecedor, não contará mais com suporte a partir de Dezembro/2017.
- A ausência de um segundo storage para replicação/backup dos dados cria um ponto único de falha para a operação da nossa rede, a despeito dos demais elementos do DataCenter da FEEC proverem redundância.

Consideramos que uma alternativa viável a curto prazo é a aquisição de um novo storage, com características de desempenho mais atuais, o qual seria dedicado a armazenar as versões primárias dos dados de usuário e discos virtuais, ficando o equipamento atual para replicação (online) destes dados e para backup. Posto que mais da metade do espaço de armazenamento é dedicado hoje para backup, a configuração inicial do novo storage não precisaria incluir tanta capacidade de armazenamento quanto o atual.

Por fim, há que se considerar que a alternativa de se realizar esta replicação em algum ambiente fora da FEEC é interessante e inclusive mais robusta, mas deve-se ter em mente que no momento a configuração do backbone do campus não permite implementar facilmente esta solução.