Evaluating the Impact of Undetected Disk Errors in RAID Systems

Apresentado por: Ivo Lopes e Pedro Pinho

Índice

Introdução

- Apresentam um modelo de falhas UDE (Undetected Disk Errors), e uma framework para simular erros em grande escala.
- A framework permite modelar sistemas de armazenamento arbitrários, cargas de trabalho e estimar a taxa de erros UDE.
- Apesar da crescente fiabilidade dos sistemas de armazenamento actual, estudos recentes mostram que existe uma nova classe de falhas, as falhas UDE que representam um desafio também crescente com o aumento das capacidades de armazenamento.

Introdução (II)

- As falhas UDE são falhas que o disco rígido não detecta. Apôs uma ordem de de escrita ou de leitura, dá-se um evento mecânico que ira causar a falha e o disco não se apercebe da situação , relatando a operação como bem sucedida.
- Estas falhas são particularmente preocupantes em sistemas RAID, já que estes são utilizados em grandes sistemas de armazenamento, e tem como intuito proteger os dados através de redundância, combinando um ou mais discos.

Introdução (III)

- Um sistema RAID pode funcionar de diversos modos. Para uma melhor compressão apresentamos a seguinte tabela.

Nível	Descrição	Nº Discos mínimo	Eficiência de espaço	Tolerância
Raid 0	Aumento de desempenho	2	n	0
Raid 1	Tolerância a falhas	2	1	n-1
Raid 5	Tolerância a falhas e>Espaço útil	3	n -1	n-1
Raid 6	Tolerância a falhas e>Espaço útil	3	n -2	n-2

Tabela 2-Tipos de sistemas RAID

Introdução (IV)

- As falhas UDE tem duas classes distintas as UWD e URD, sendo estas Undetected Write Errors e Undetected Read Errors.
- Para estas falhas a protecção que se obtêm através do uso de sistemas RAID não é suficiente.
- Neste artigo os autores apresentam um modelo que permite simular e detectar estes eventos.

Undetected Disk Errors

- A tabela 3 apresenta o sumario dos possíveis UDE.

I/O Type	*UDE type*	*Manifestation*
Write (UWE)	Dropped Write	Stale data
	Near off-track write	Possible stale data on read
	Far off-track write	Stale data on intended track
		Corrupt data on written track
Read (URE)	Near off-track read	Possible stale data
	Far off-track read	Corrupt data

Undetected Disk Errors (II)

- Os erros UDE, acontecem quando o sistema da uma ordem de escrita ou leitura para o disco.
- Não são detectados pelo hardware nem pelo software.
- Os sistemas RAID não detectam, nem protegem contra UDE.
- Os UDE podem ser de escrita ou de leitura.
- Devem-se a falhas mecânicas, falhas de software ou de firmware.

Undetected Write Errors

- Os erros UWE, acontecem quando o sistema da uma ordem de escrita para o disco, e acontecem do seguinte modo:
  - A cabeça de escrita do disco posiciona-se correctamente mas não escreve os dados.
  - A cabeça de escrita falha o posicionamento e escreve os dados na pista errada
    - Falha “Near off-track”, ou seja falha por pouco.
    - Falha “Far off-track”, afasta-se mais da pista pretendida.
- Manifestam-se na fase de leitura sendo sistemáticos e persistentes para os blocos afectados.

Undetected Read Errors

- Os erros URE acontecem quando é dada uma ordem de leitura ao disco, e podem dar-se do seguinte modo:
  - A cabeça de leitura falha o posicionamento e lê os dados na pista errada
    - Falha “Near off-track”, ou seja falha por pouco.
    - Falha “Far off-track”, afasta-se mais da pista pretendida.
- Os erros URE manifestam-se naturalmente na fase de leitura, e são ocasionais, podendo dar-se uma vez num determinado bloco, e logo a seguir já não de darem nesse mesmo bloco.

Detecção

- Os sistemas de armazenamento implementados raramente tem mecanismos para detectar UDE´s.
- Esta demonstrado que nos sistemas RAID, o data scrubing por si só, não chega para proteger contra todos os UDE´s.
  - Data scrubing é o método mais comum usado nos sistemas RAID para a detecção e correcção de erros, consiste no varrimento em intervalos de tempo definidos de todos os blocos de todos os discos, a procura de erros, e corrigindo-os caso existam.

Detecção (II)

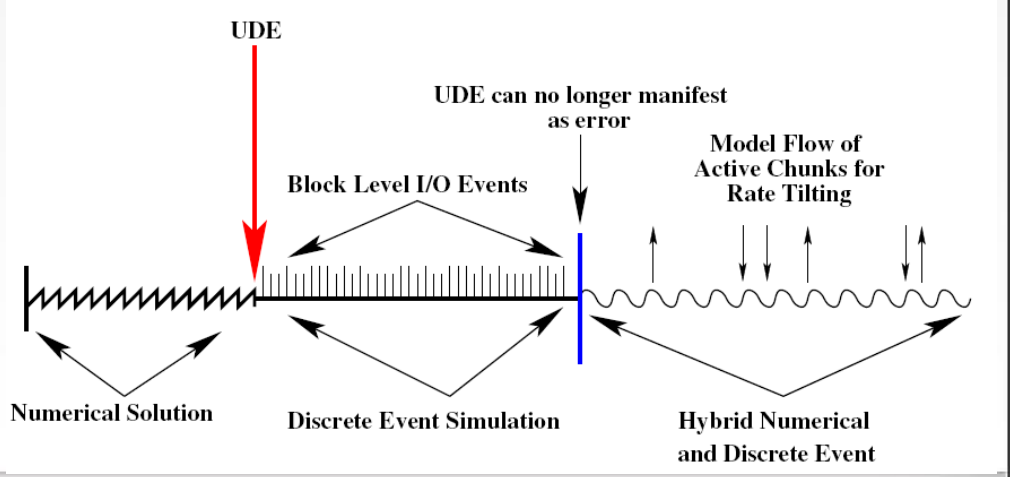
- Existem contudo propostas para detecção deste tipo de erros, neste paper o modelo de detecção foca-se no “data parity appendix method”.
- Este modelo baseia-se na adição de um numero de sequencia a cada bloco escrito, que será o mesmo em cada ciclo de escrita.
- Os UDE´s podem então ser detectados, comparando os números de sequencia lidos dos blocos com os números de sequencia armazenados nos bits de paridade. Se a comparação não for igual, o erro UDE é detectado.

Framework (I)

- Apresentam uma framework híbrida para simular o armazenamento em grande escala;
- Combina eventos de simulação discreta em múltiplos níveis com uma analise numérica num modelo híbrido;
- O metodo tira vantagem da relação com uma stream I/O para temporariamente mudar entre os métodos numéricos, Numerical Solution, Discrete Event Simulation e Hybrid Numerical and Discrete Event.

Framework (II)

- A Framework combina 3 modos de operação:
  - Numerical Solution
  - Discrete Event Simulation
  - Hybrid Numerical and Discrete Event



Framework (III)

- Esta técnica consegue ser mais eficiente a nível de tempo e espaço do que apenas as abordagens até então existentes que se baseavam apenas no Discrete Event Simulation;
- A framework permite novas implementações por parte do utilizador; ex: novas técnicas de detecção.

Framework (IV)

- A framework permite dinamicamente mudar entre os 3 modos, para assim maximizar a eficiência enquanto mantém a eficácia na detecção de dados corruptos;
- Os UDE estão altamente relacionados com a carga de trabalho do disco e a framework foi pensada com base no mesmo.

Resultados

- Para este paper foram estudados 3 cenários diferentes, para apurar os efeitos dos UDEs:
- - Large-scale storage system
- - Large entreprise storage system
- - Small business storage system

Resultados (II)

- O Large-scale storage system, é composto por 1000 discos, cada um com capacidade para 1 terabyte
- O Large entreprise storage system é composto por 512 discos, cada um com capacidade para 300 gigabytes
- O Small business storage system é composto por 32 discos, cada um com capacidade para 300 gigabytes
- Todos os cenários foram testados com as 3 etapas da Framework

Resultados (III)

System			
Mitigation	**Large**	**Enterprise**	**Small Business**
NO	0.718	0.718	0.718
YES	0.0028	0.0028	0.0028

Tabela 4-Proporçao de UDEs por tamanho do sistema.

Resultados (IV)

System			
Mitigation	**Abstract**	**Read Heavy**	**Write Heavy**
NO	0.718	0.275	0.887
YES	0.0028	0.0011	0.0035

Tabela 5-Proporçao de UDEs por carga do sistema.

Resultados (V)

Estimated rate for various rates of UDEs/IO							
		10^{-11}		10^{-12}		10^{-13}	
System	**Mitigation**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
Large scale	No	$6.278 \cdot 10^{-7}$	$6.281 \cdot 10^{-10}$	$6.282 \cdot 10^{-8}$	$7.430 \cdot 10^{-11}$	$6.282 \cdot 10^{-9}$	$6.324 \cdot 10^{-12}$
Large scale	Yes	$2.415 \cdot 10^{-9}$	$6.312 \cdot 10^{-11}$	$2.466 \cdot 10^{-10}$	$6.502 \cdot 10^{-12}$	$2.519 \cdot 10^{-11}$	$5.705 \cdot 10^{-13}$
Enterprise	No	$3.217 \cdot 10^{-7}$	$3.813 \cdot 10^{-10}$	$3.218 \cdot 10^{-8}$	$3.813 \cdot 10^{-11}$	$3.221 \cdot 10^{-9}$	$2.195 \cdot 10^{-12}$
Enterprise	Yes	$1.259 \cdot 10^{-9}$	$2.503 \cdot 10^{-11}$	$1.262 \cdot 10^{-10}$	$4.042 \cdot 10^{-12}$	$1.253 \cdot 10^{-11}$	$2.213 \cdot 10^{-13}$
Small	No	$2.012 \cdot 10^{-8}$	$1.595 \cdot 10^{-11}$	$2.012 \cdot 10^{-9}$	$1.110 \cdot 10^{-12}$	$2.012 \cdot 10^{-10}$	$1.589 \cdot 10^{-13}$
Small	Yes	$7.930 \cdot 10^{-11}$	$1.633 \cdot 10^{-12}$	$7.868 \cdot 10^{-12}$	$1.602 \cdot 10^{-13}$	$7.857 \cdot 10^{-13}$	$2.201 \cdot 10^{-14}$

Tabela 6-Media e desvio padrao de UDEs por tamanho do sistema.

Resultados (VI)

Estimated rate for various rates of UDEs/IO							
		10^{-11}		10^{-12}		10^{-13}	
System	**Mitigation**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
Abstract	No	$6.278 \cdot 10^{-7}$	$6.281 \cdot 10^{-10}$	$6.282 \cdot 10^{-8}$	$7.430 \cdot 10^{-11}$	$6.282 \cdot 10^{-9}$	$6.324 \cdot 10^{-12}$
Abstract	Yes	$2.415 \cdot 10^{-9}$	$6.312 \cdot 10^{-11}$	$2.466 \cdot 10^{-10}$	$6.502 \cdot 10^{-12}$	$2.519 \cdot 10^{-11}$	$5.705 \cdot 10^{-13}$
Read Heavy	No	$2.404 \cdot 10^{-7}$	$2.465 \cdot 10^{-10}$	$2.405 \cdot 10^{-8}$	$4.692 \cdot 10^{-11}$	$2.401 \cdot 10^{-9}$	$3.436 \cdot 10^{-12}$
Read Heavy	Yes	$9.345 \cdot 10^{-10}$	$2.526 \cdot 10^{-11}$	$9.310 \cdot 10^{-11}$	$2.195 \cdot 10^{-12}$	$9.476 \cdot 10^{-12}$	$3.231 \cdot 10^{-13}$
Write Heavy	No	$7.764 \cdot 10^{-7}$	$1.061 \cdot 10^{-9}$	$7.763 \cdot 10^{-8}$	$9.004 \cdot 10^{-11}$	$7.766 \cdot 10^{-9}$	$6.756 \cdot 10^{-12}$
Write Heavy	Yes	$3.048 \cdot 10^{-9}$	$4.592 \cdot 10^{-11}$	$3.014 \cdot 10^{-10}$	$3.902 \cdot 10^{-12}$	$3.038 \cdot 10^{-11}$	$7.858 \cdot 10^{-13}$

Tabela 7-Media e desvio padrao de UDEs por carga do sistema.

Conclusões

- Os autores consideram que os UDE´s irão ter uma importância cada vez mais significativa a medida que os sistemas de armazenamento vão crescendo, algo que parece ser comprovado pelos resultados obtidos.
- Este problema torna-se relevante mesmo em sistemas de armazenamento de pequena escala usados em pequenos negócios.
- Sem o uso de medidas de contenção a taxa de UDE´s pode disparar mesmo antes de se atingir o MTBF dos sistemas usados.

Conclusões (II)

Mean Interval				
System	**Mitigation**	10^{-11}	10^{-12}	10^{-13}
Large	No	18.43 days	184.2 days	5.04 years
Large	Yes	13.13 years	128.6 years	1258 years
Ent.	No	35.98 days	0.9854 years	9.844 years
Ent.	Yes	25.19 years	251.3 years	2531 years
Small	No	1.576 years	15.76 years	157.6 years
Small	Yes	399.9 years	4030 years	40358 years

Tabela 8-Intervalo medio de UDE´s para carga mista

Conclusões (III)

- Os resultados demonstram que efectuar “data scrubing” semanalmente não chega já que o “scrubing” pode precisar de áreas onde se deram UDE´s para reconstruir dados noutras áreas, o que vai levar a falhas graves no sistema.
- Os resultados mostram também que o uso de discos talhados para o utilizador comum, faz com que o tempo médio entre em falhas para os sistemas de maior dimensão , seja inferior a um ano.
- Dado o crescente uso destas drives em sistemas RAID, torna-se evidente as limitações destes sistemas.

Conclusões (IV)

- Os autores afirmam no entanto que o uso de técnicas complementares como o uso de números de sequencia, podem conter de modo eficaz os UDE´s durante o MTBF dos discos destinados ao consumidor geral.
- Tal método de contenção ira duplicar as operações de leitura, mas vai resultar num decréscimo da taxa de erros numa ordem de magnitude de 2 a 3 graus.

Fim