Universidade do Minho

Mestrado em Engenharia Informática

Engenharia dos Sistemas de Computação - 2014/2015

Testes de Desempenho E/S (IOzone)

5 de Maio de 2015

Resumo

Este trabalho demonstra o resultado do Benchmark IOzone em vários nós do SeARCH no Departamento de Informática da Universidade do Minho. Desta forma será possível analisar os recursos de armazenamento com os testes intensivos que o IOzone providencia.

Fábio Gomes pg27752

Índice

Introdução

O problema que nos foi apresentado consiste em analisar a performance dos discos do cluster. A ferramenta utilizada é o IOzone é um Benchmark para sistemas de ficheiros. Gera e mede uma variedade de operações sobre ficheiros, utilizando inclusive *mmap* e POSIX threads. Vencendo até o *Infoworld Bossie Awards* em 2007 como a melhor ferramenta de E/S para ficheiros.

Com estes dados recolhidos foi possível gerar gráficos comparativos entre todos os discos testados e fazer conclusões.

Caracterização do Sistema

Cada nó tem um disco associado, na tabela seguinte faço a sua associação. Para obter os Discos de cada nó recorri ao programa `tentakel`, para dispersão, e `udevadm`, para listagem, com a seguinte parametrização:

```
tentakel -g compute_linux "/sbin/udevadm info -a -p /sys/class/block/sda/sda5 -q env | grep MODEL"
```

Para referências futuras apenas será mencionado o nome do Disco como identificação.

Nó do Cluster	**Disco**
431-3	MB0500EBNCR
431-5	SAMSUNG_HD502HI
431-6	SAMSUNG_HD502HJ
432-1	MM0500EBKAE
541-1	WDC_WD10EZRX-00A8LB0
641-8	INTEL_SSDSC2BW240A3F
641-19	INTEL_SSDSC2BW240A4
652-1	WDC_WD20NPVT-00Z2TT0
662-6	INTEL_SSDSC2BW120A4

IOZONE e Parâmetros de teste

Para este trabalho utilizei o `IOzone` versão `3.397`, escolhi output para ficheiro binário *excel* e nome específico do ficheiro temporário para não coincidir com outros testes concorrentes. Cada ficheiro output e temporário tem o nome do nó associado, neste caso o *431-3*.

```
/opt/iozone/bin/iozone -Ra -b 431_3.xls -f 431_3.tmp
```

Testes utilizados

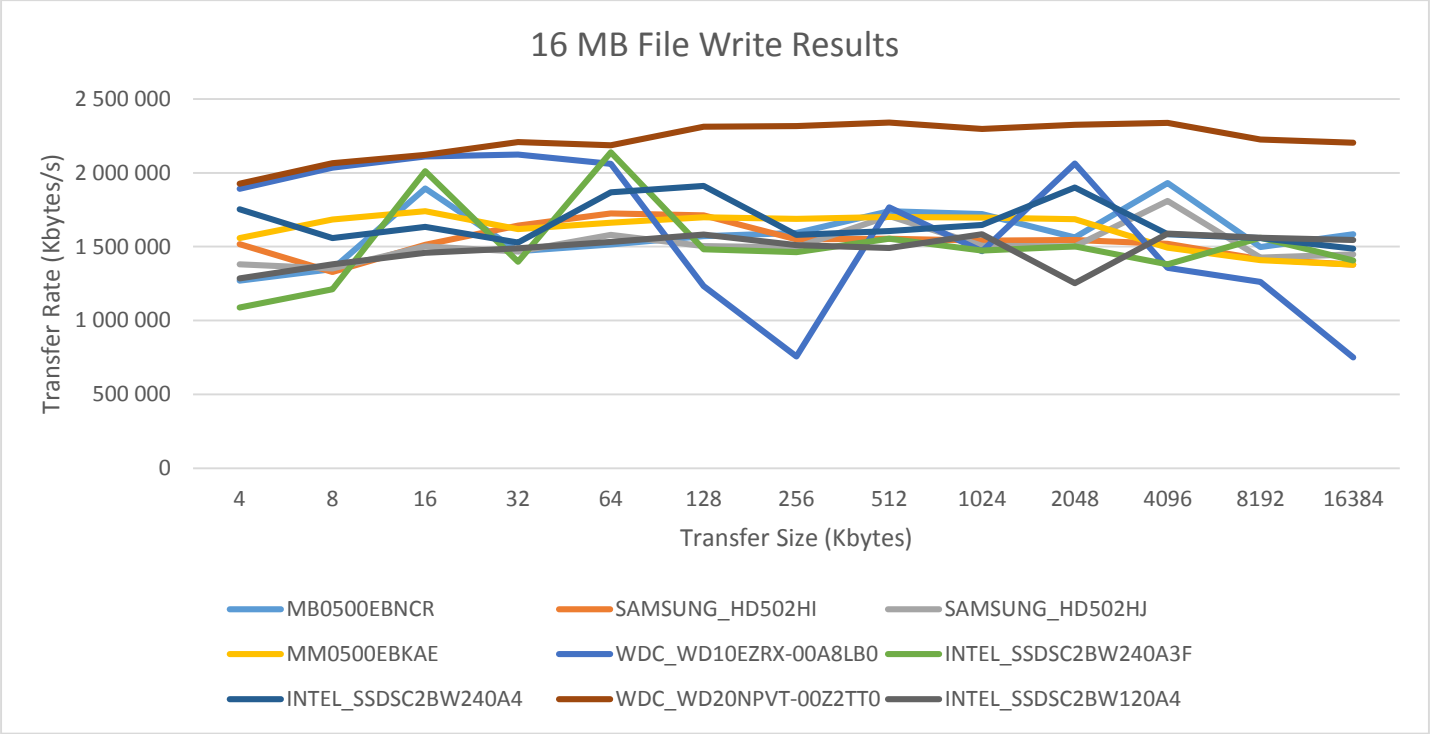
Realizei 13 testes para 512 MB de tamanho máximo de ficheiro. Sendo os testes os seguintes:

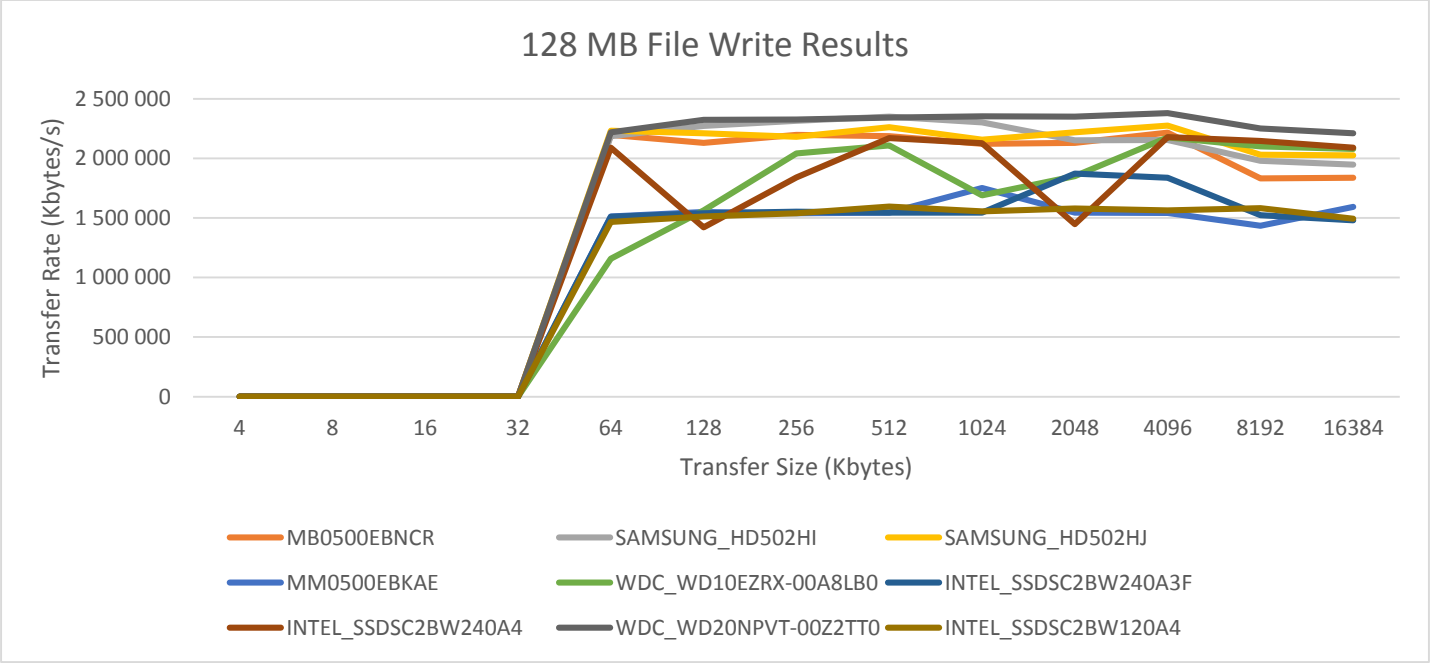
1. Write
2. Rewrite
3. Read
4. Reread
5. Random read
6. Random write
7. Backward read
8. Record rewrite
9. Stride read
10. Fwrite
11. Frewrite
12. Fread
13. Freread.

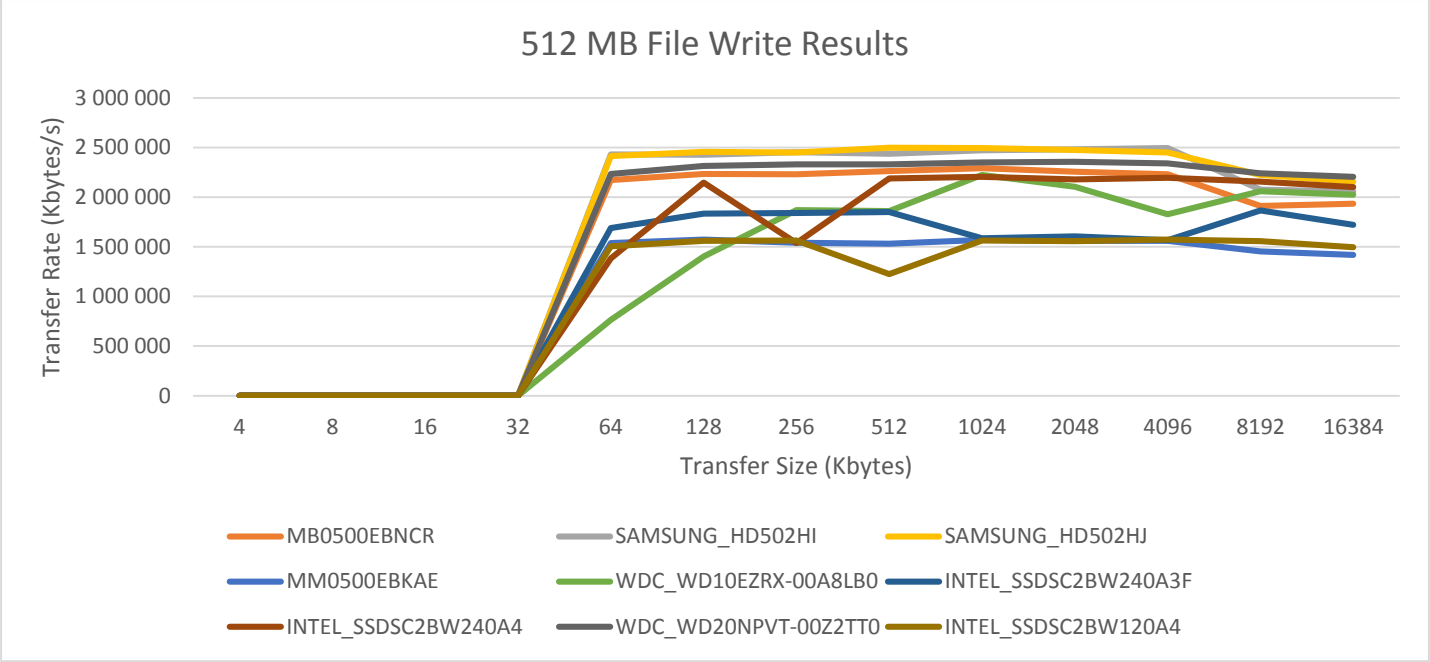
Write

Este teste mede a performance de escrita num ficheiro. Quando um novo ficheiro é escrito não são só os dados guardados mas também a informação sobre onde eles são guardados no disco, chamados de *metadados*. Consiste na informação de diretoria, espaço alocado e outras informações associadas ao ficheiro que não fazem parte dos dados do ficheiro em si. É normal que no início deste teste a performance seja mais baixa devido a esta informação extra.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB de ficheiro de teste.







Conclusão

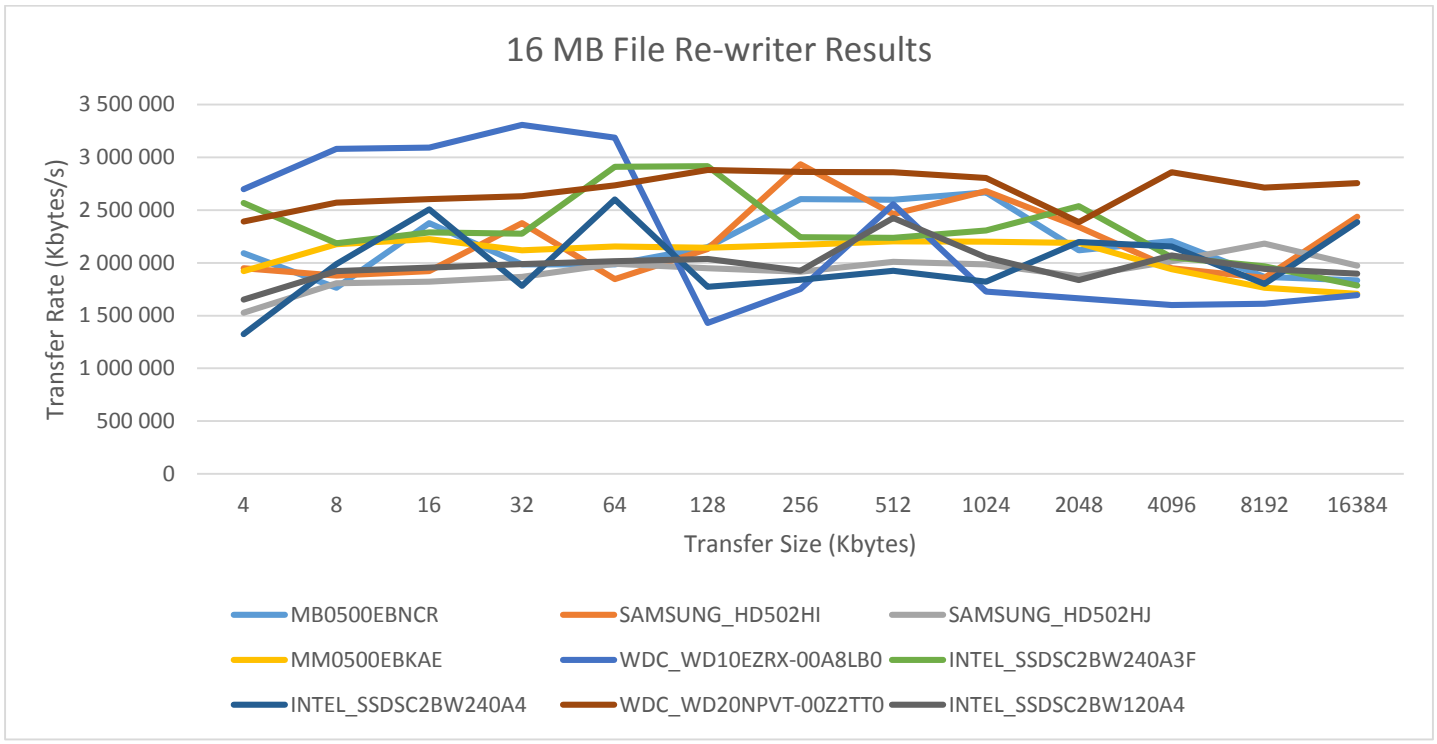
Para 16 MB e 128 MB o **WDC_WD20NPVT-00Z2TT0** foi o melhor, ficando em segundo lugar no teste de 512 MB quando perde para o **SAMSUNG_HD502HJ**.

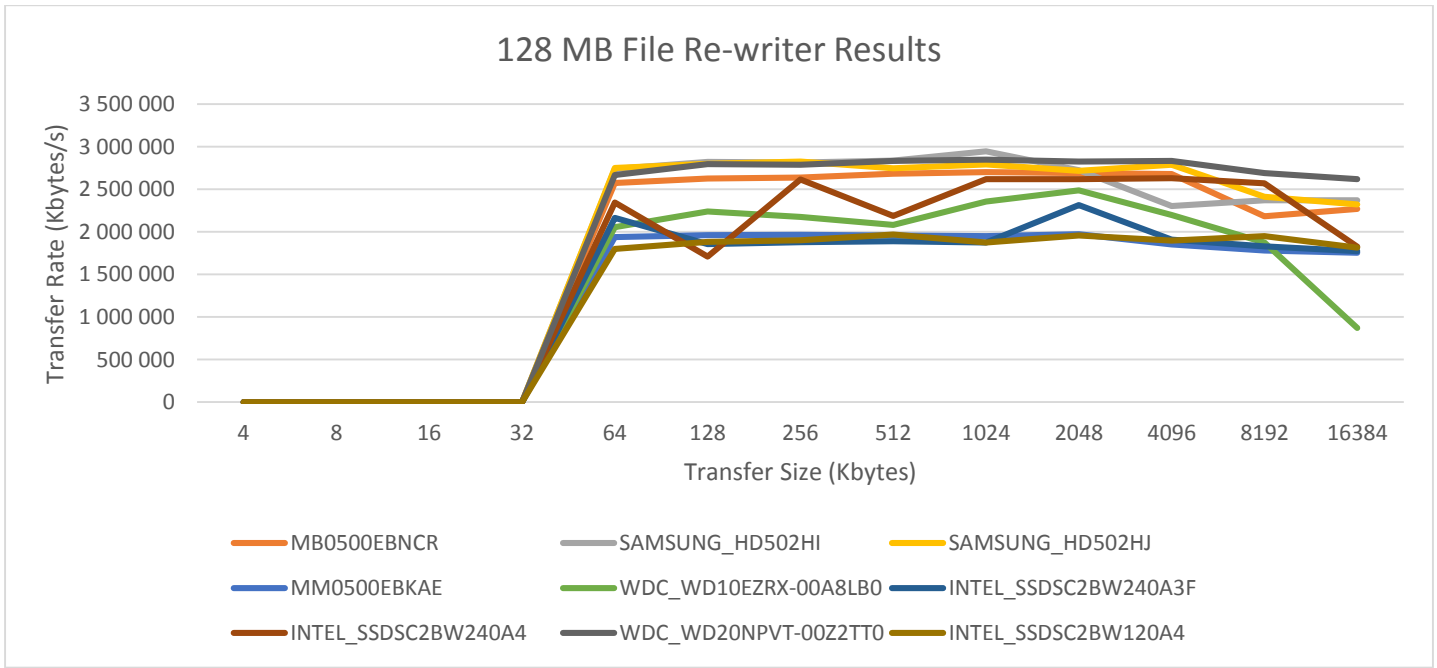
No geral o **WDC_WD20NPVT-00Z2TT0** obteve os melhores resultados.

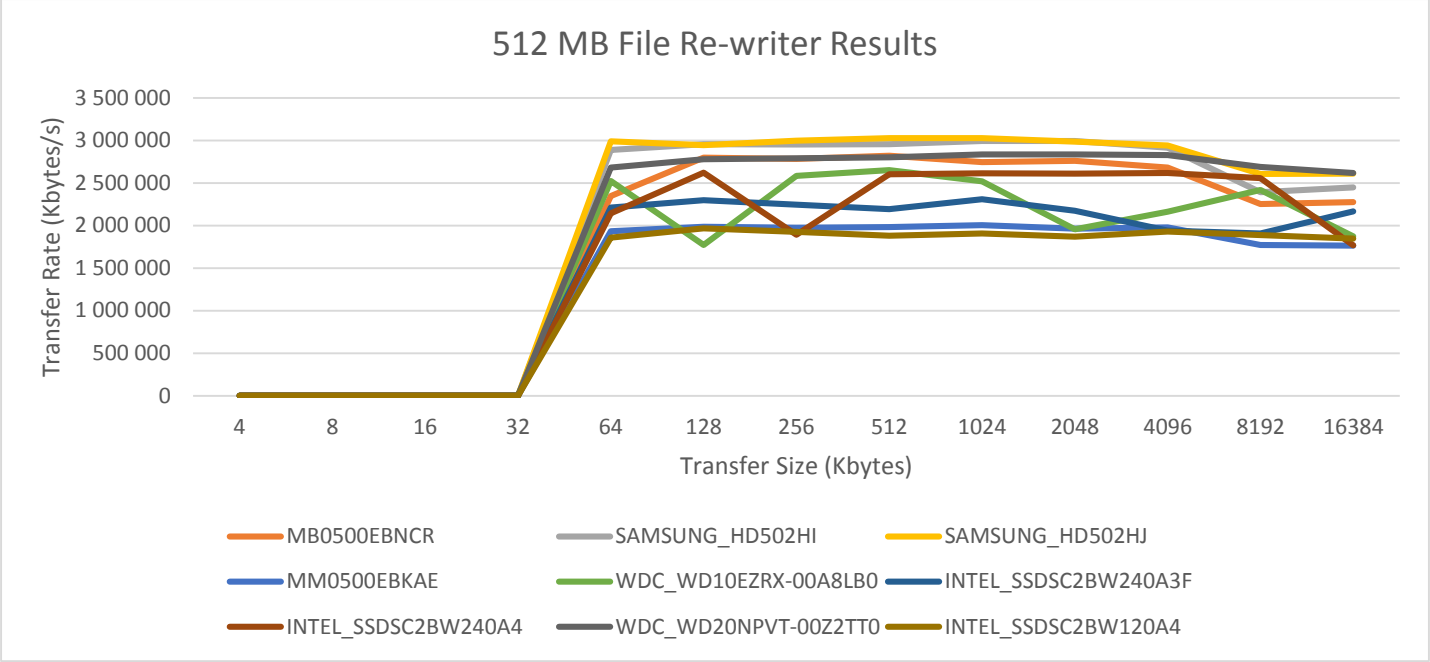
Re-Write

Este teste mede a performance de escrita de um ficheiro que já existe. Sendo assim é preciso menos trabalho pois já existem metadados guardados. É expectável que a performance deste teste seja superior ao anterior, pelas razões apresentadas.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







Conclusão

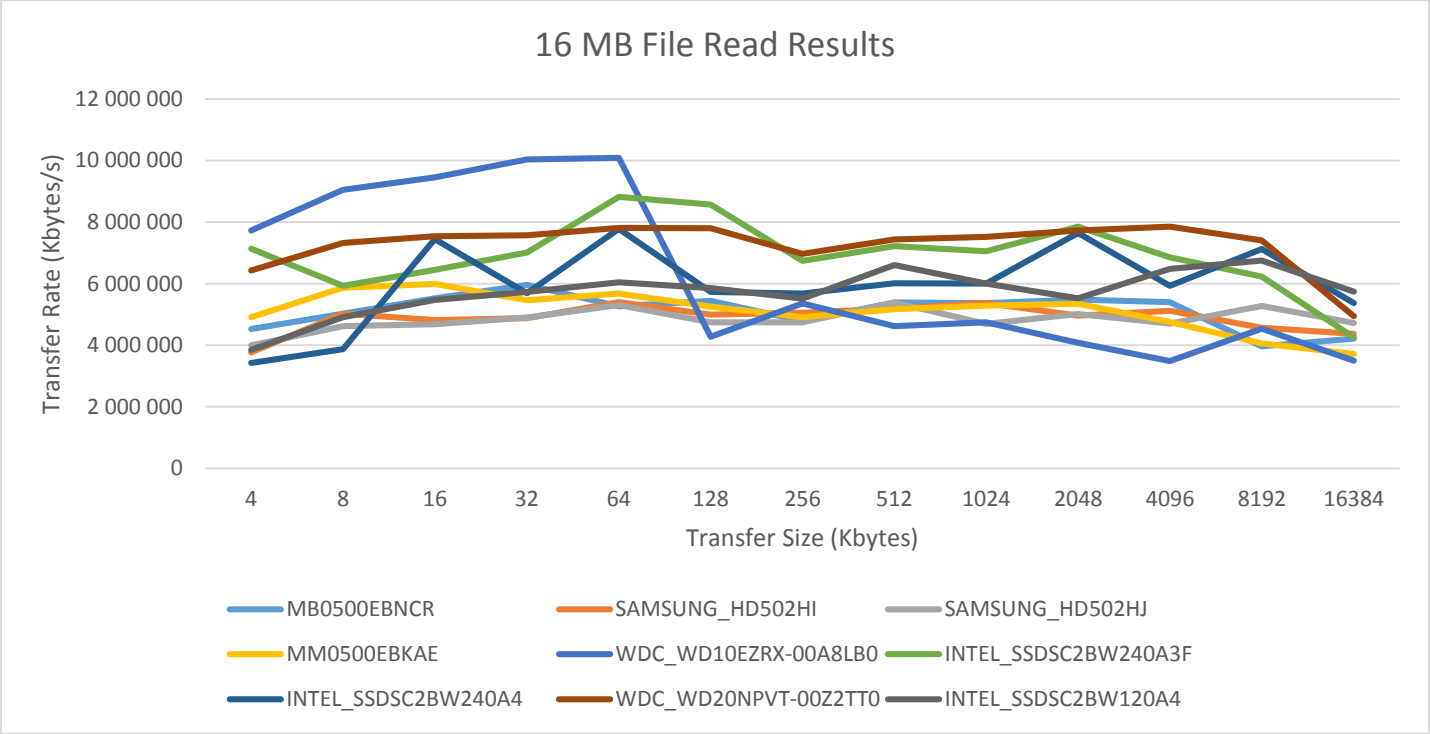
Como previsto todos os testes foram melhores que os de Write.

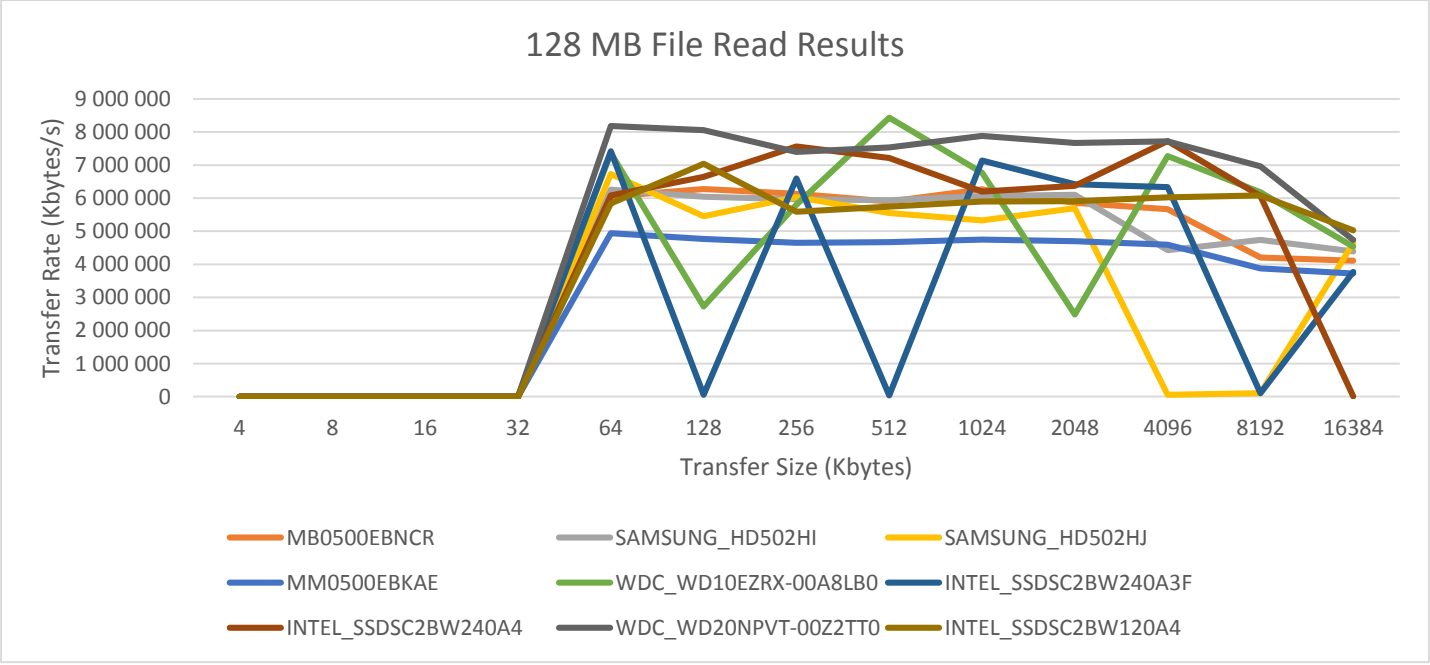
Para 16 MB o **WDC_WD20NPVT-00Z2TT0** foi o melhor, depois para 128 MB e 512 MB há de novo pouca diferença entre o **WDC_WD20NPVT-00Z2TT0** e o **SAMSUNG_HD502HJ**.

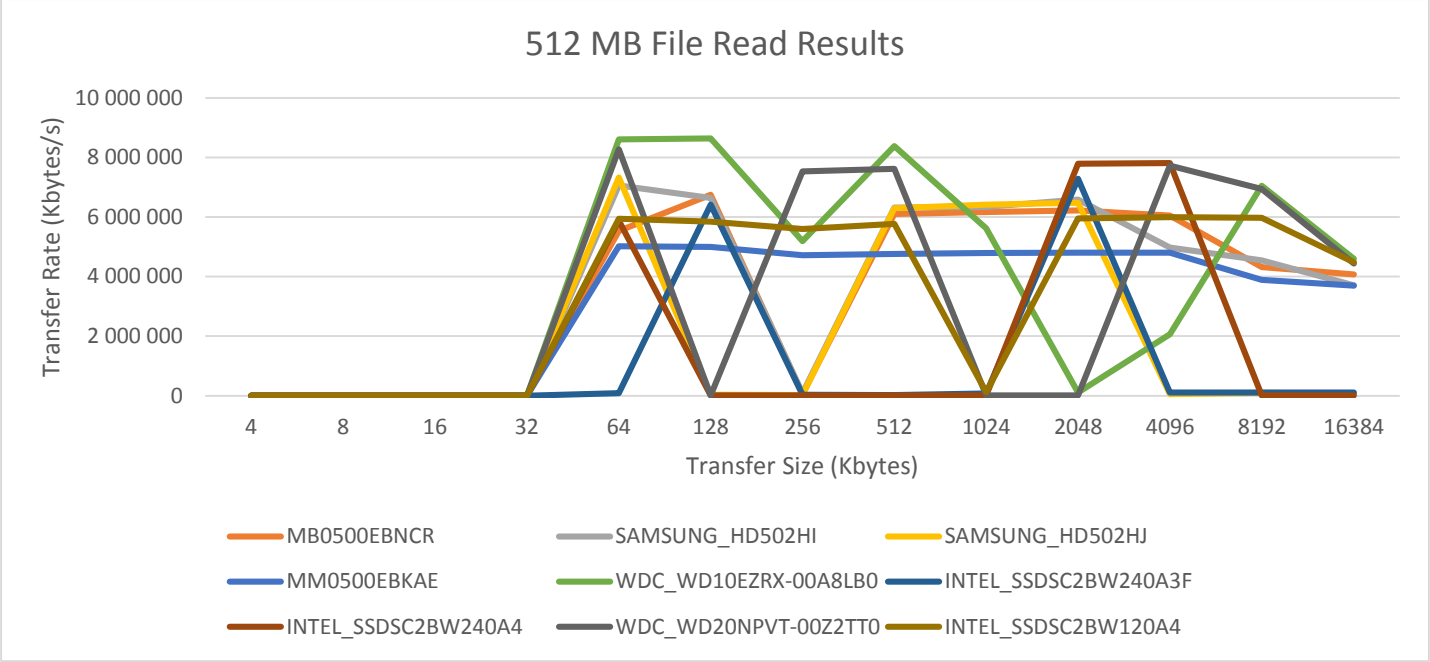
No geral o **WDC_WD20NPVT-00Z2TT0** obteve os melhores resultados.

Read

Este teste mede a performance de leitura de um ficheiro existente.
Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







Conclusão

Há medida que iam aumentando os valores do ficheiro de teste as oscilações foram aumentando.

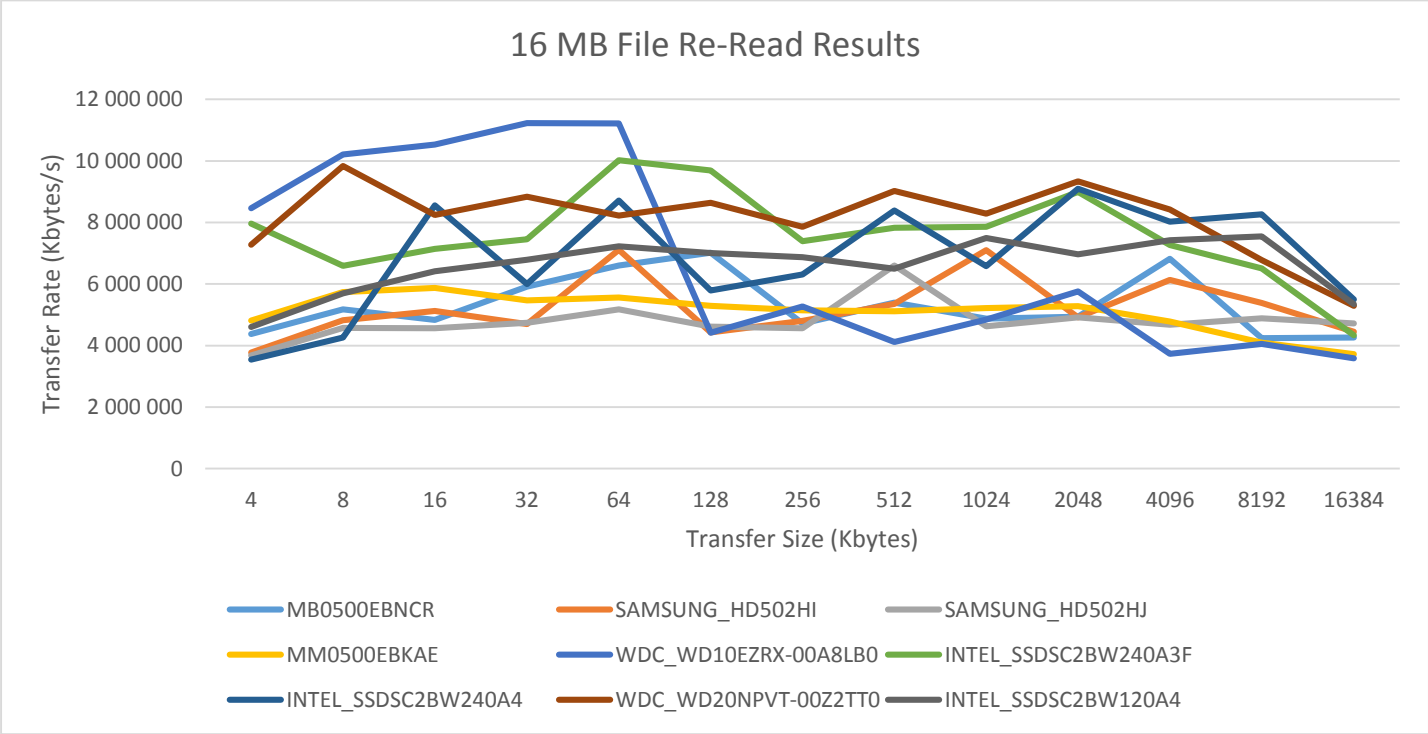
Para 16 MB e 128 MB o **WDC_WD20NPVT-00Z2TT0** foi o melhor e mais consistente. Para 512 MB apenas há um disco que se revelou constante, o **MM0500EBKAE**, que também no de 126 MB também se manteve.

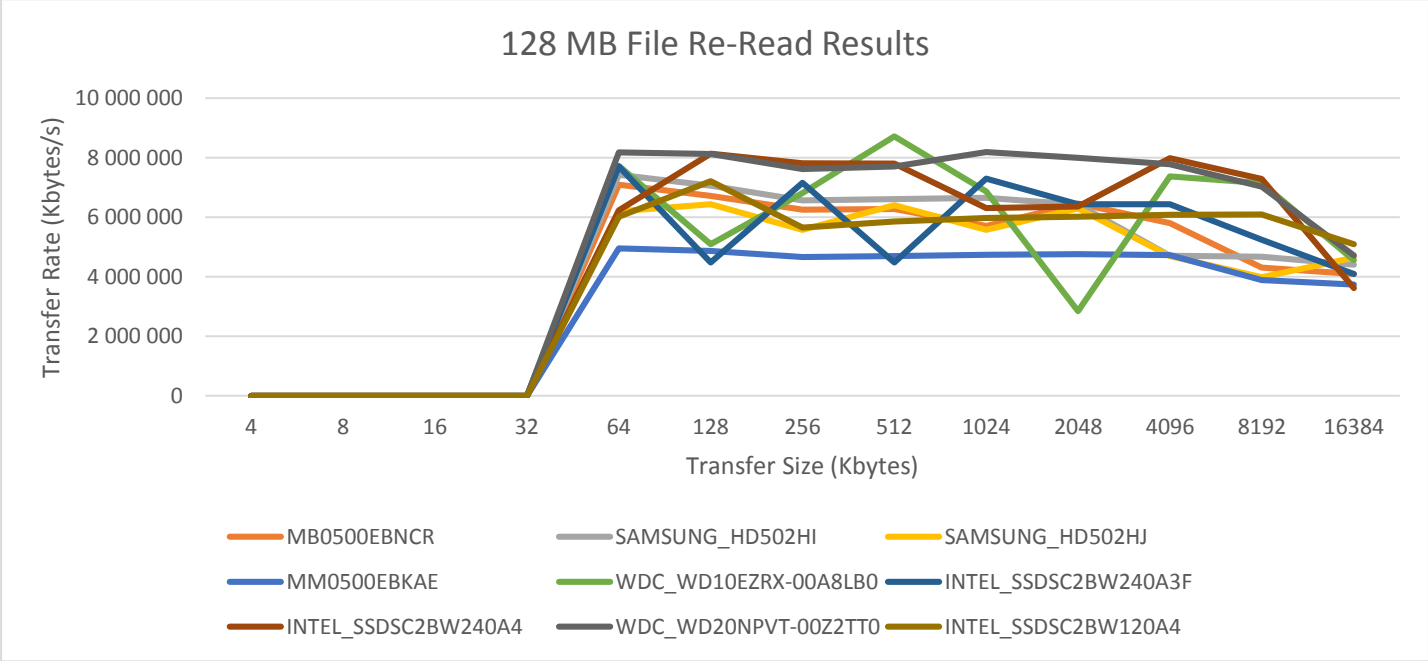
No geral o **WDC_WD20NPVT-00Z2TT0** obteve os melhores resultados.

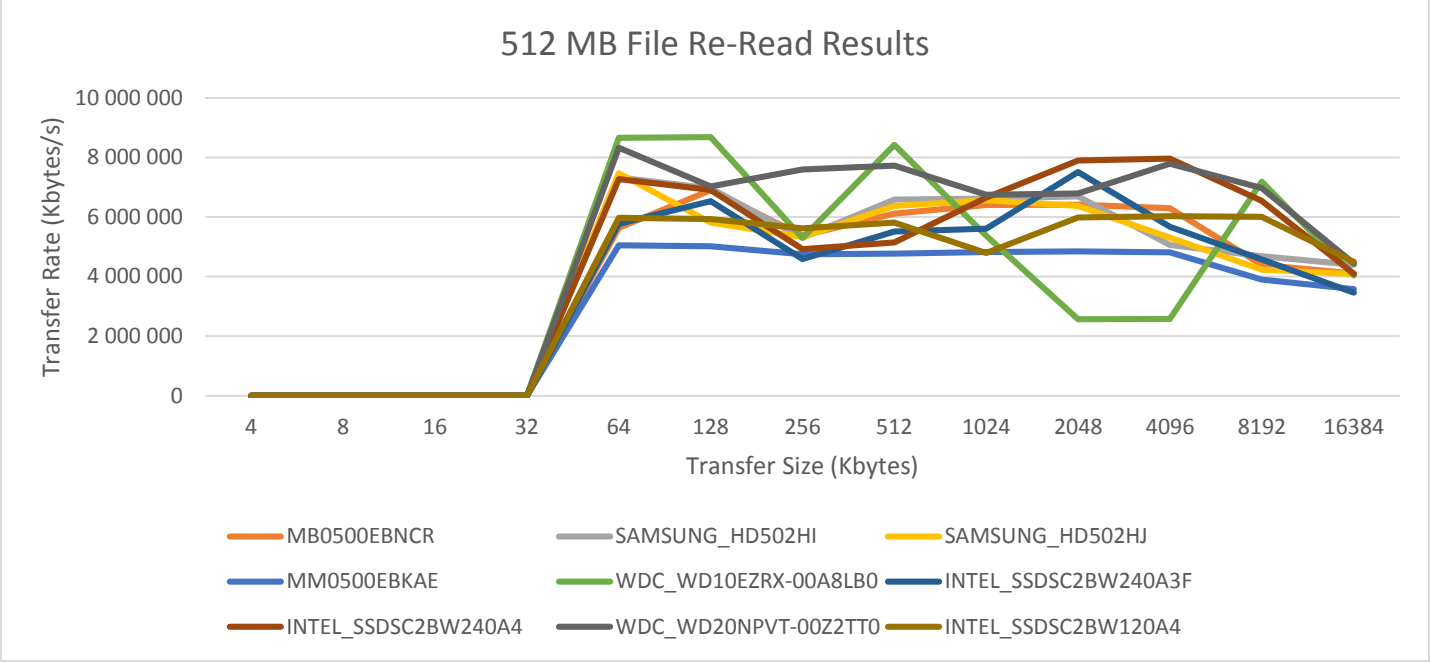
Re-Read

Este teste analisa a performance de leitura de um ficheiro que foi lido recentemente, sendo expectável que a performance seja maior que a simples leitura (Read). As caches e suas latências têm influência neste teste pois como estamos a tratar de lados que foram lidos há pouco tempo só fará sentido, para este teste, se estiverem em cache.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







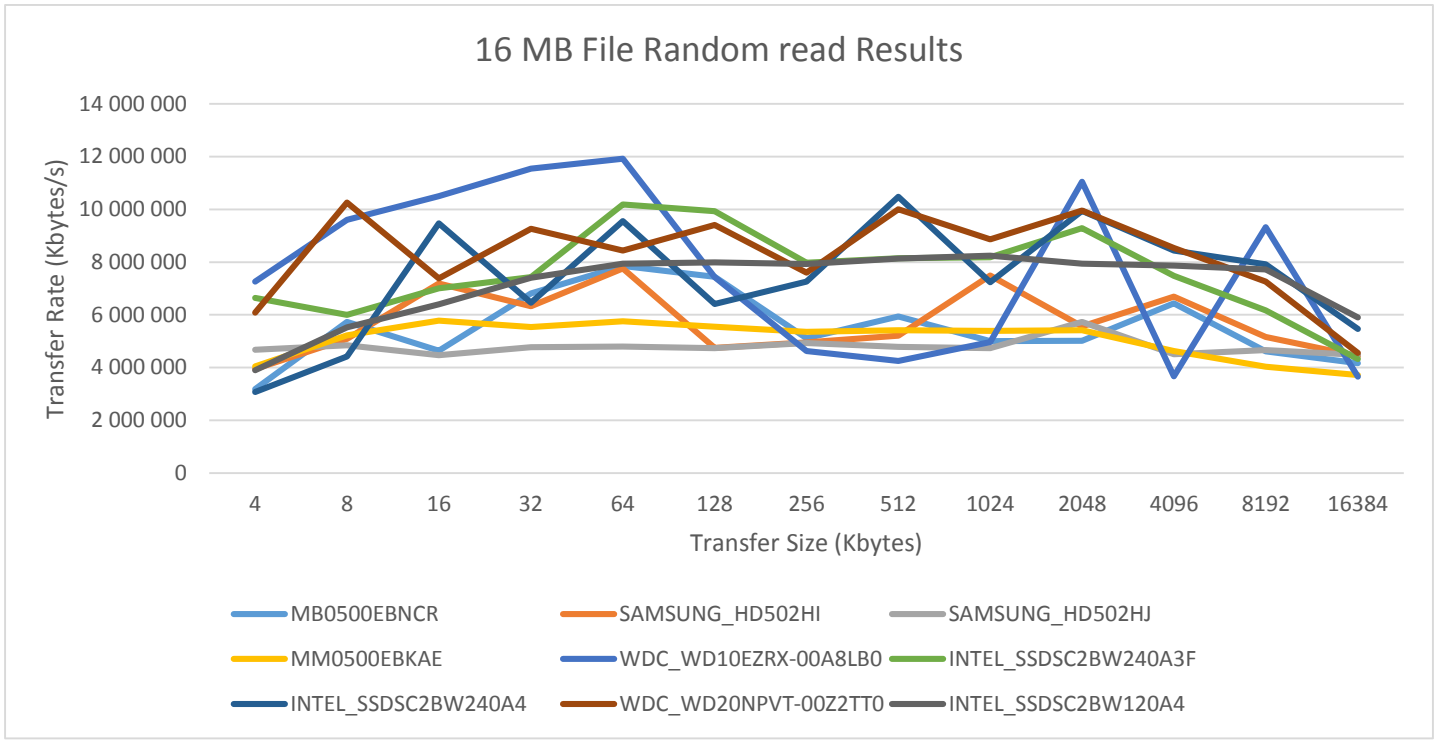
Conclusão

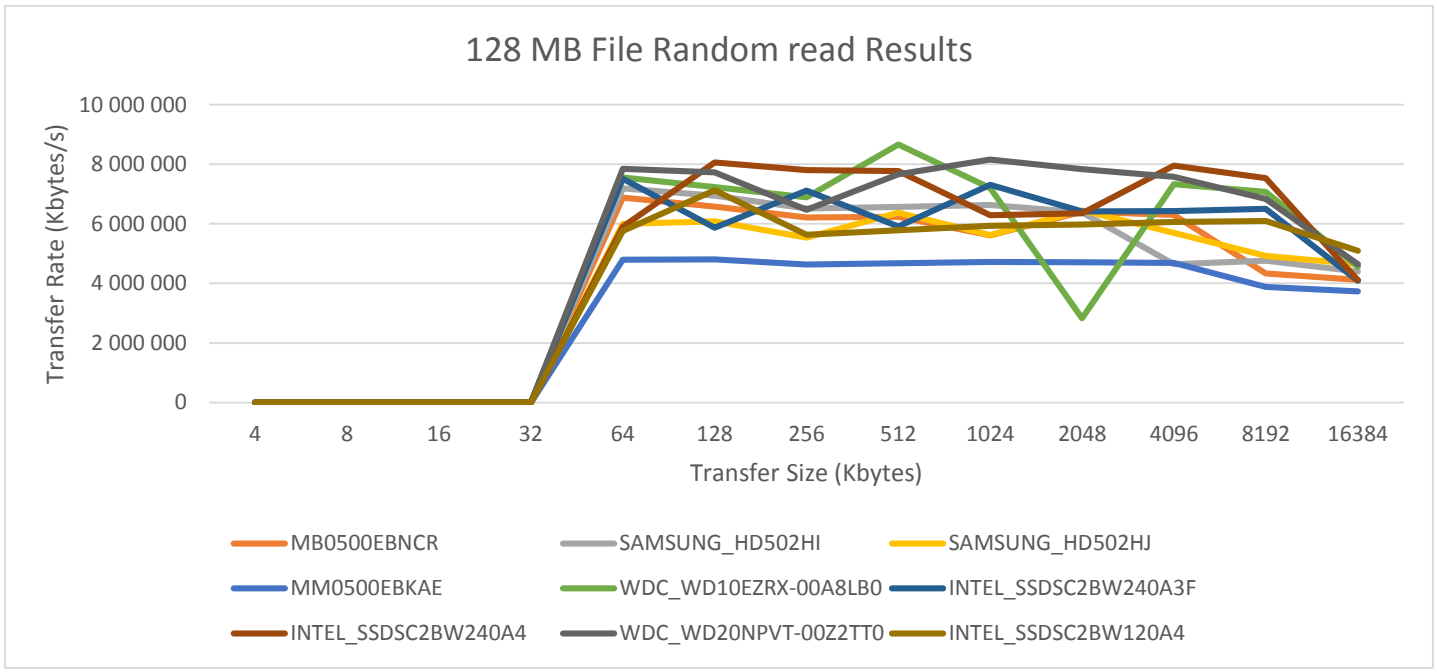
Os testes não foram muito melhores que os de Read, mas nota-se um pequeno aumento.
Nos 3 tamanhos o **WDC_WD20NPVT-00Z2TT0** foi no geral o melhor.

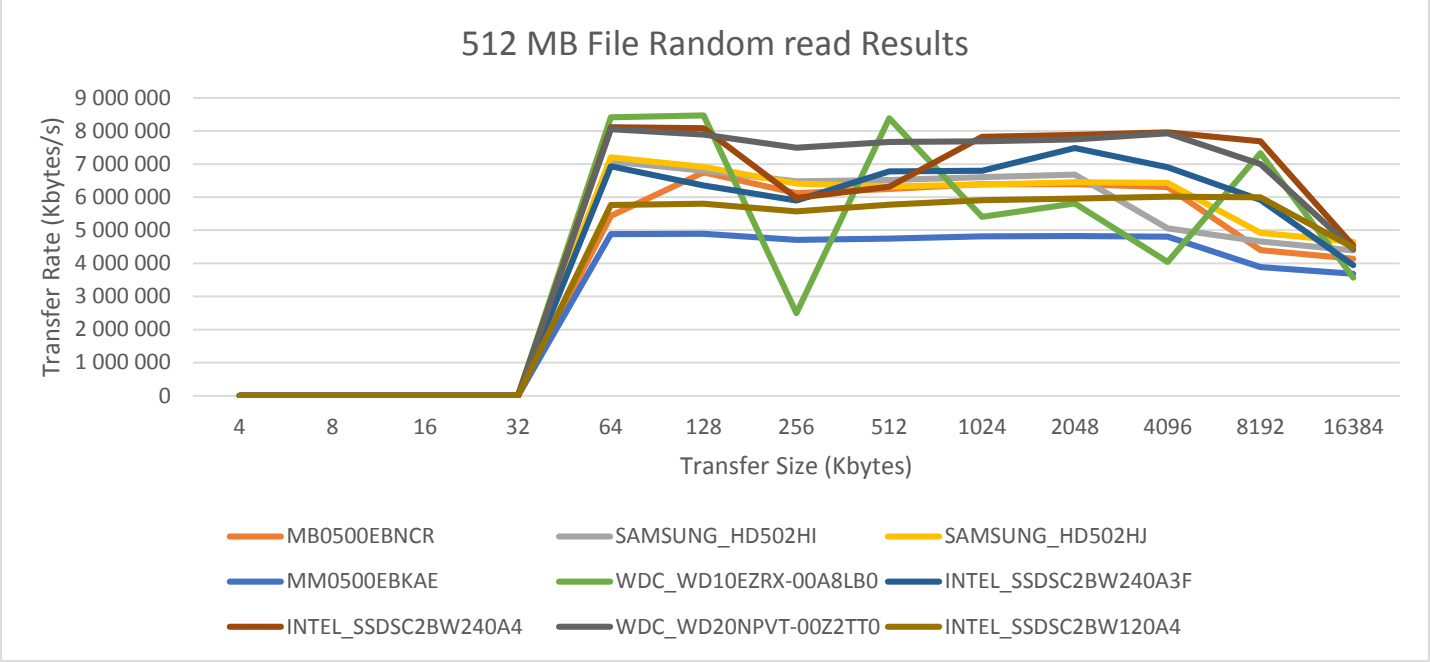
Random Read

Até agora os acessos a um ficheiro eram praticamente contínuos, mas com este teste serão aleatórios. Portanto a cache será um fator que poderá ajudar em certas ocasiões.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







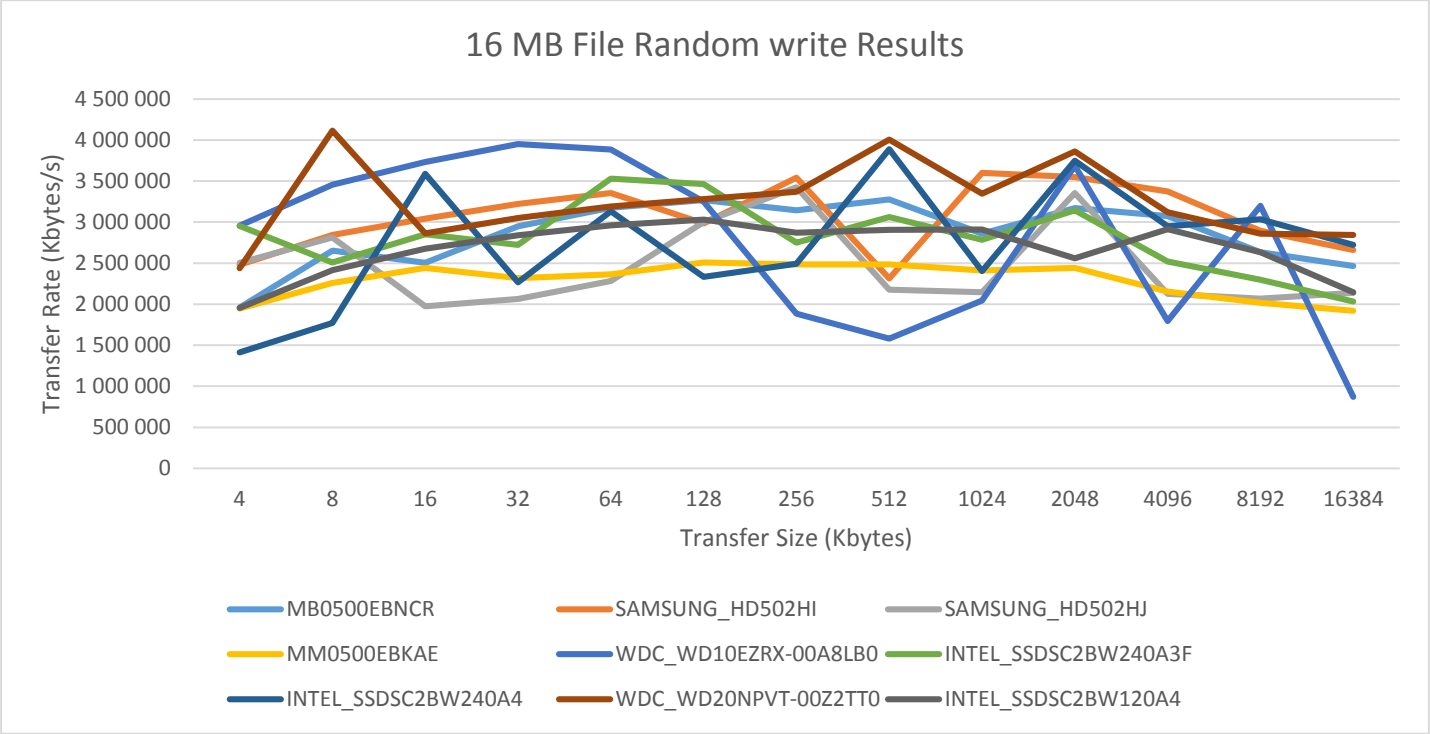
Conclusão

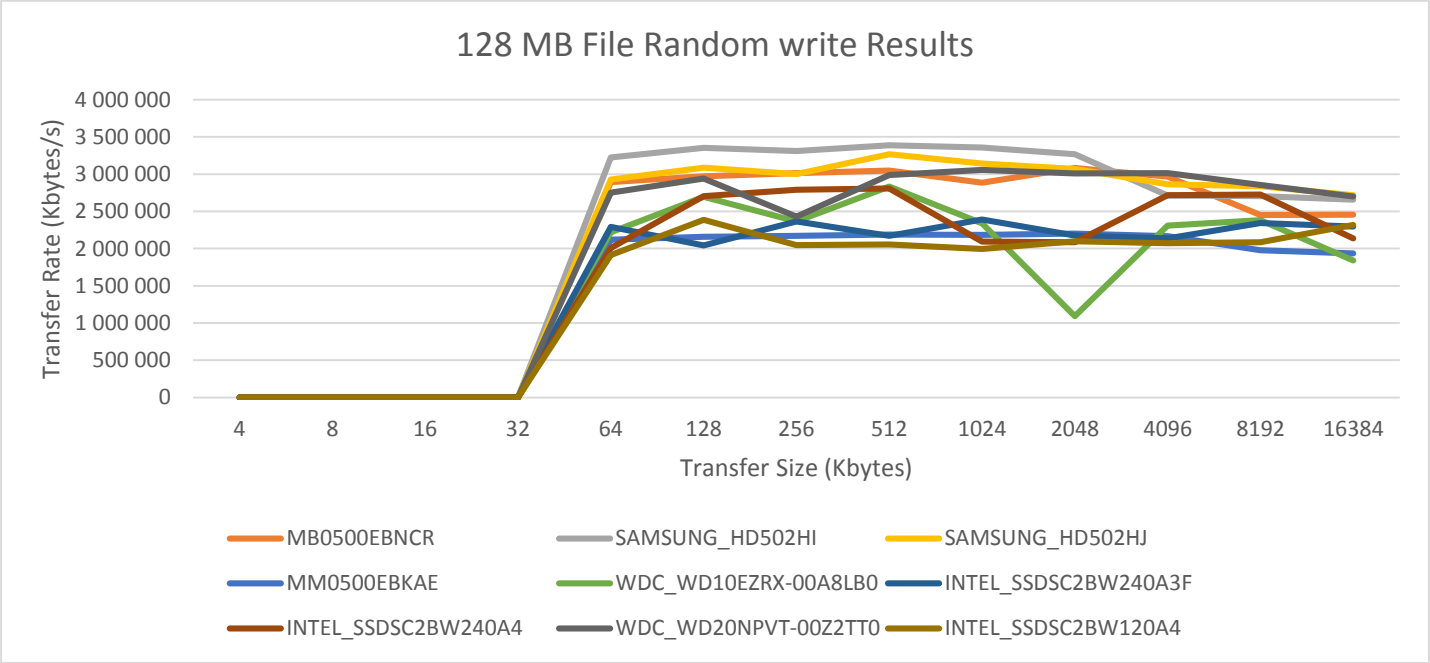
As oscilações eram previsíveis pois os acessos são irregulares.

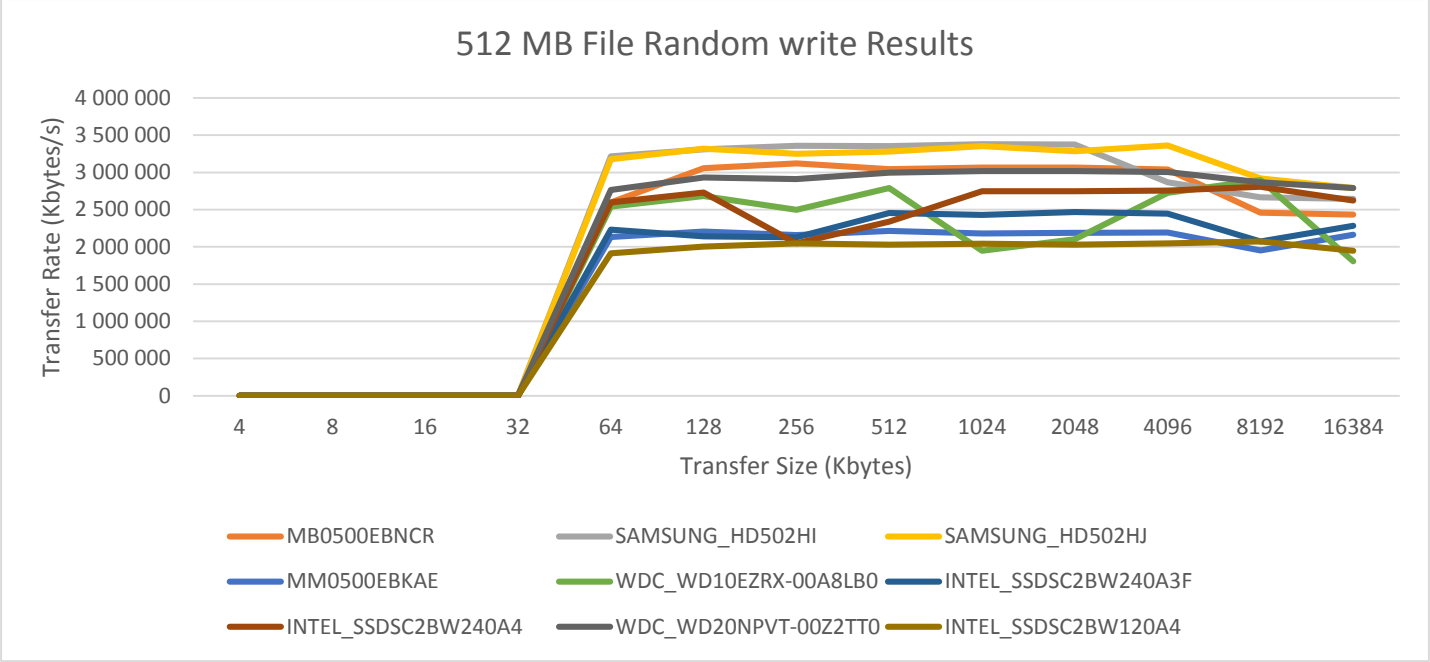
Nos 3 tamanhos o **WDC_WD20NPVT-00Z2TT0** foi no geral o melhor. Sendo que nos 128 MB o **INTEL_SSDSC2BW240A4** esteve perto do topo.

Random Write

Como o teste anterior, os acessos são aleatórios mas neste teste tratam-se de escritas.
Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







Conclusão

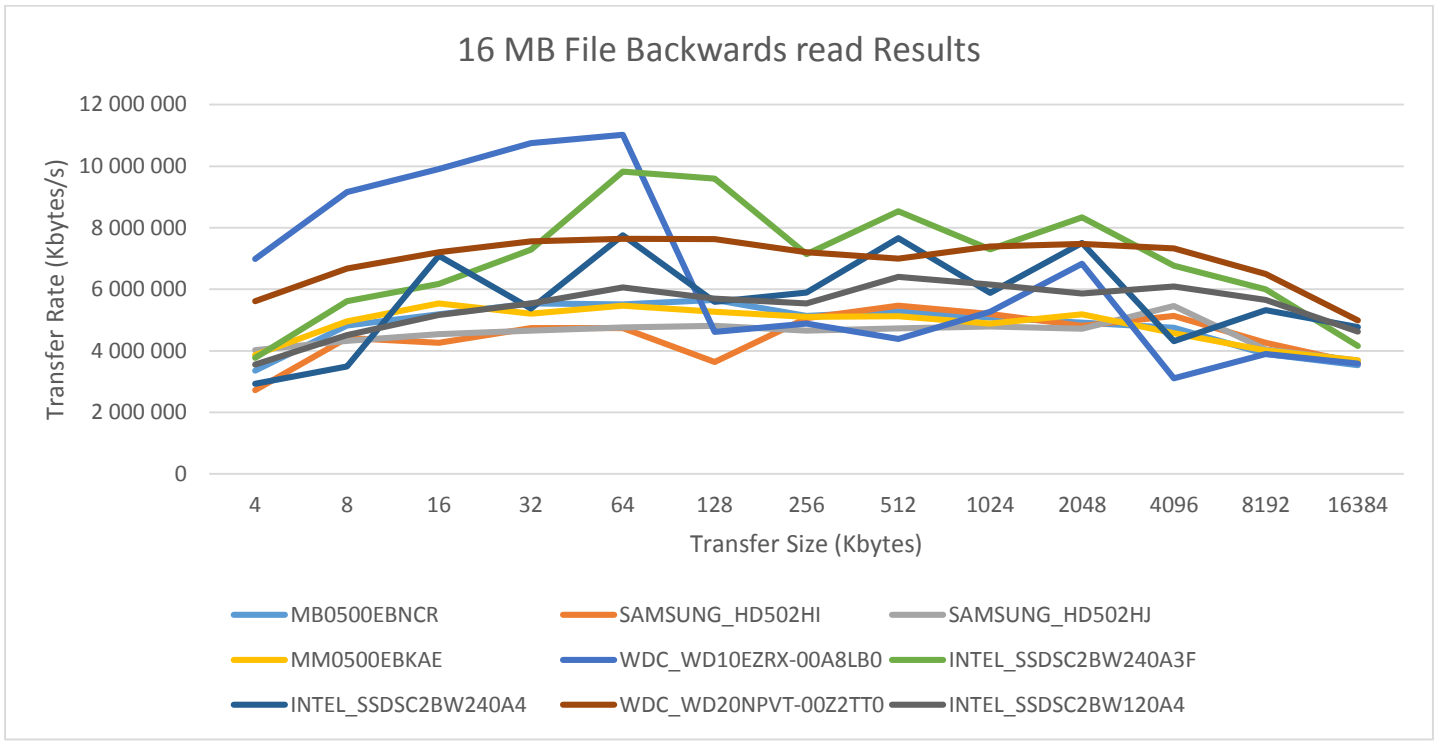
Os valores obtidos estão mais altos que os de Write normal.

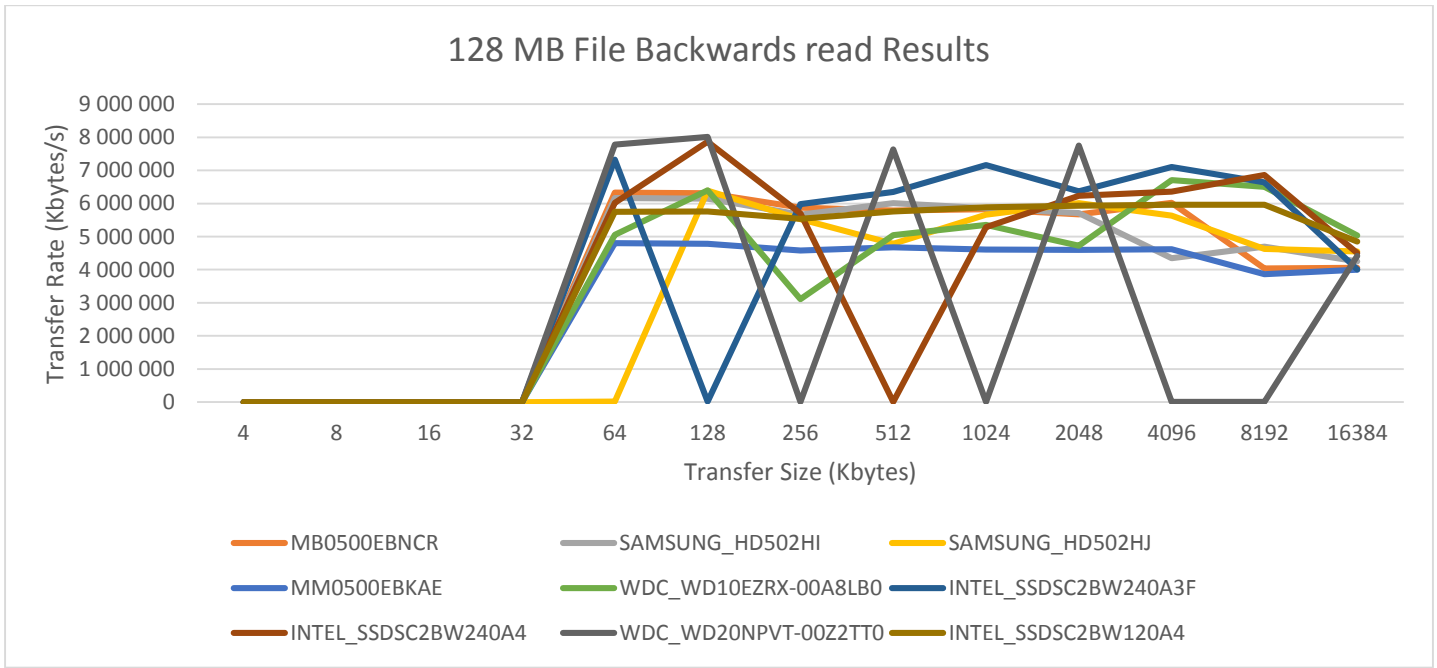
Nos 2 tamanhos iniciais o **WDC_WD20NPVT-00Z2TT0** foi no geral o melhor mas para 512 MB o **SAMSUNG_HD502HJ** leva uma ligeira vantagem.

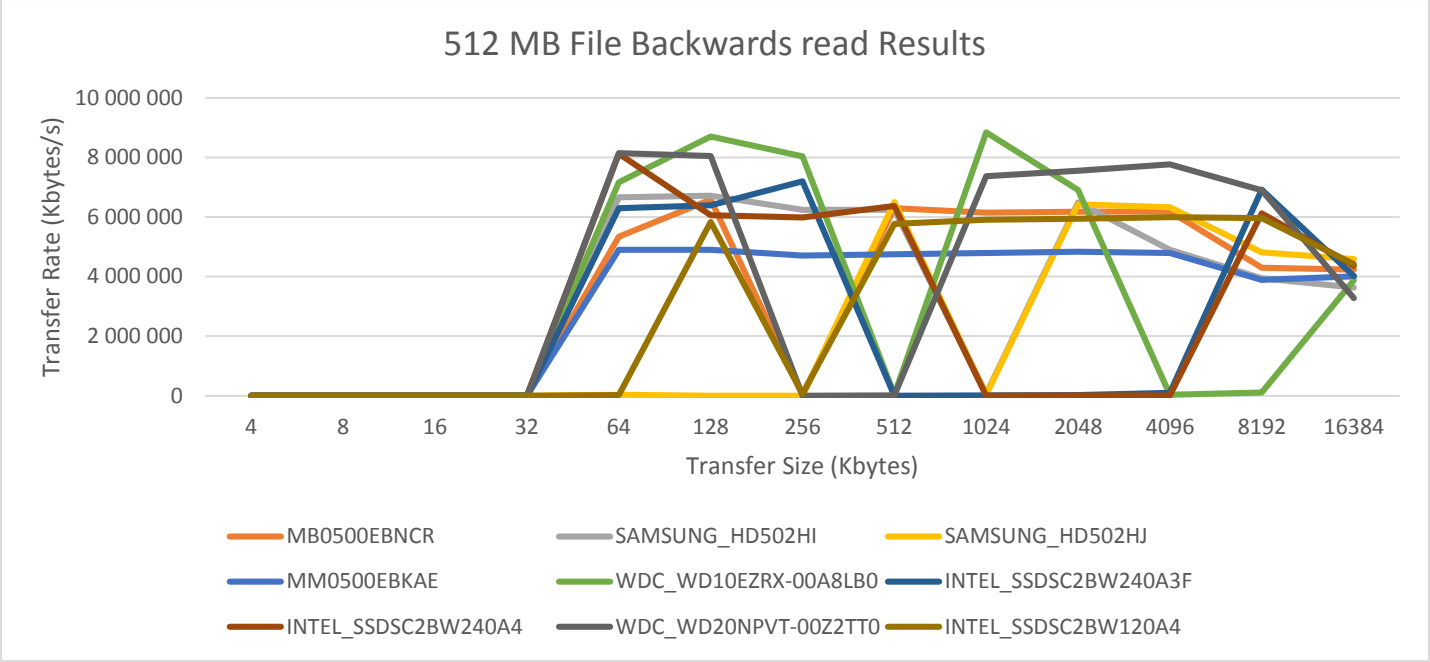
Backwards Read

Este teste irá testar a leitura de um ficheiro para trás. É uma forma estranha de ler um ficheiro mas pode acontecer que um programa tenha necessidade de o fazer. Alguns Sistemas Operativos detetam este tipo de leitura e podem aumentar a performance da leitura para trás.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







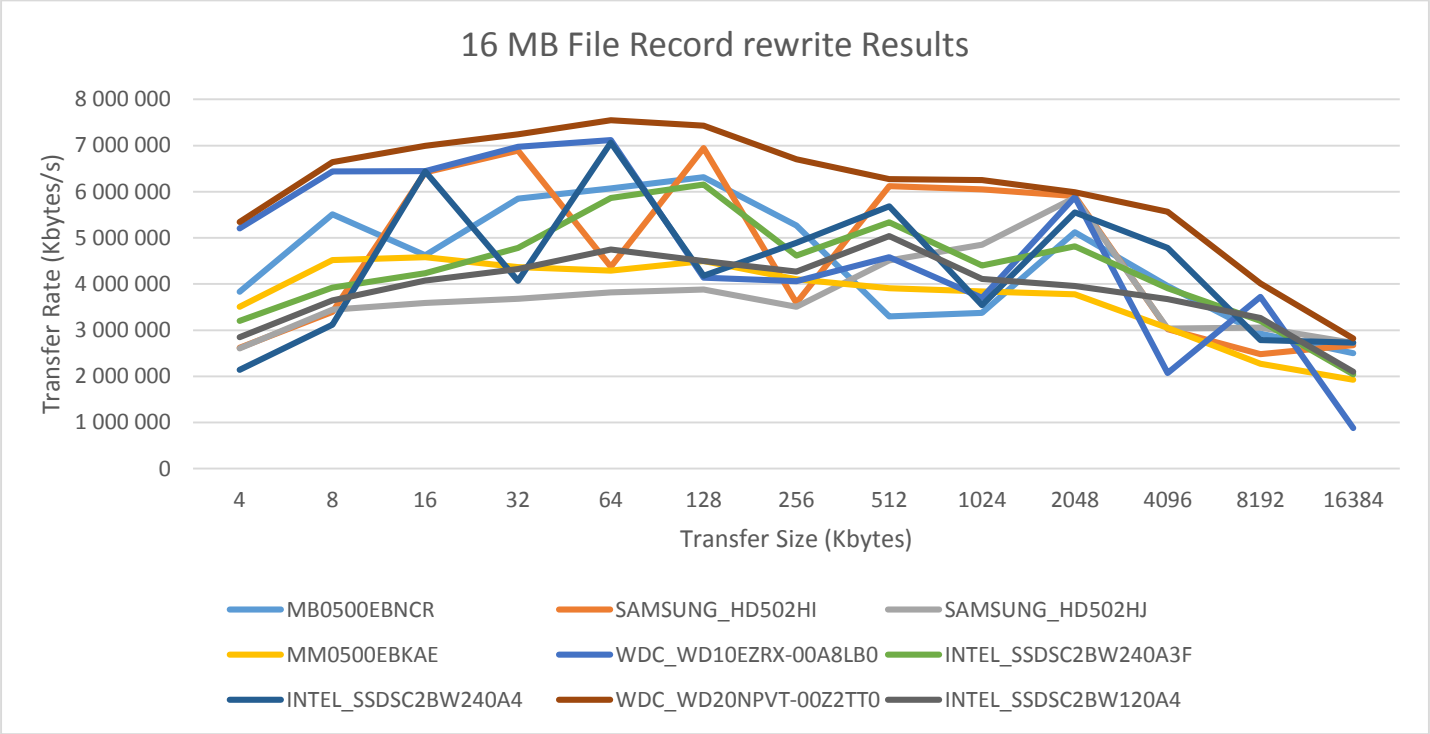
Conclusão

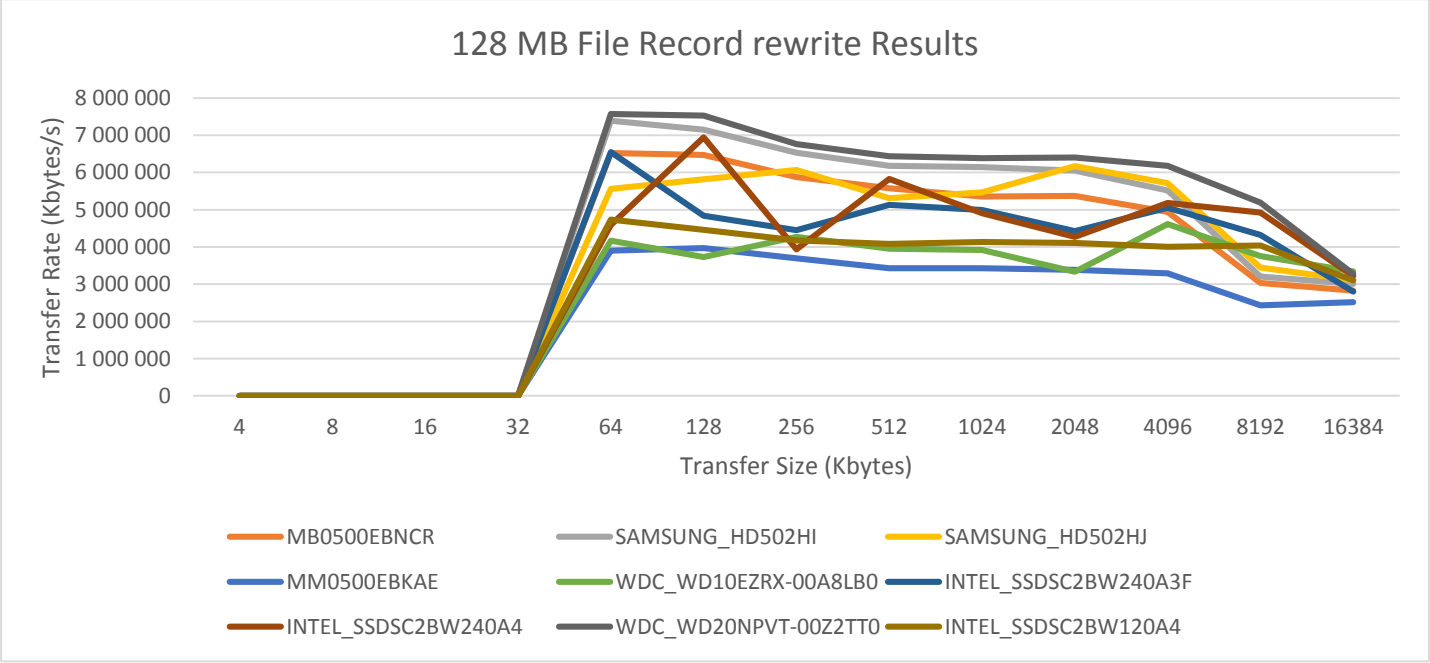
Para 16 MB o **INTEL_SSDSC2BW240A3F** apresenta melhores resultados, mas o **WDC_WD20NPVT-00Z2TT0** foi o que apresentou valores mais altos e constantes. Para 128MB o **WDC_WD10EZRX-00A8LB0** conseguiu valores altos e constantes sem oscilações bruscas. Já para 512MB o **INTEL_SSDSC2BW240A4** foi o que no total obteve maio *throughput*.

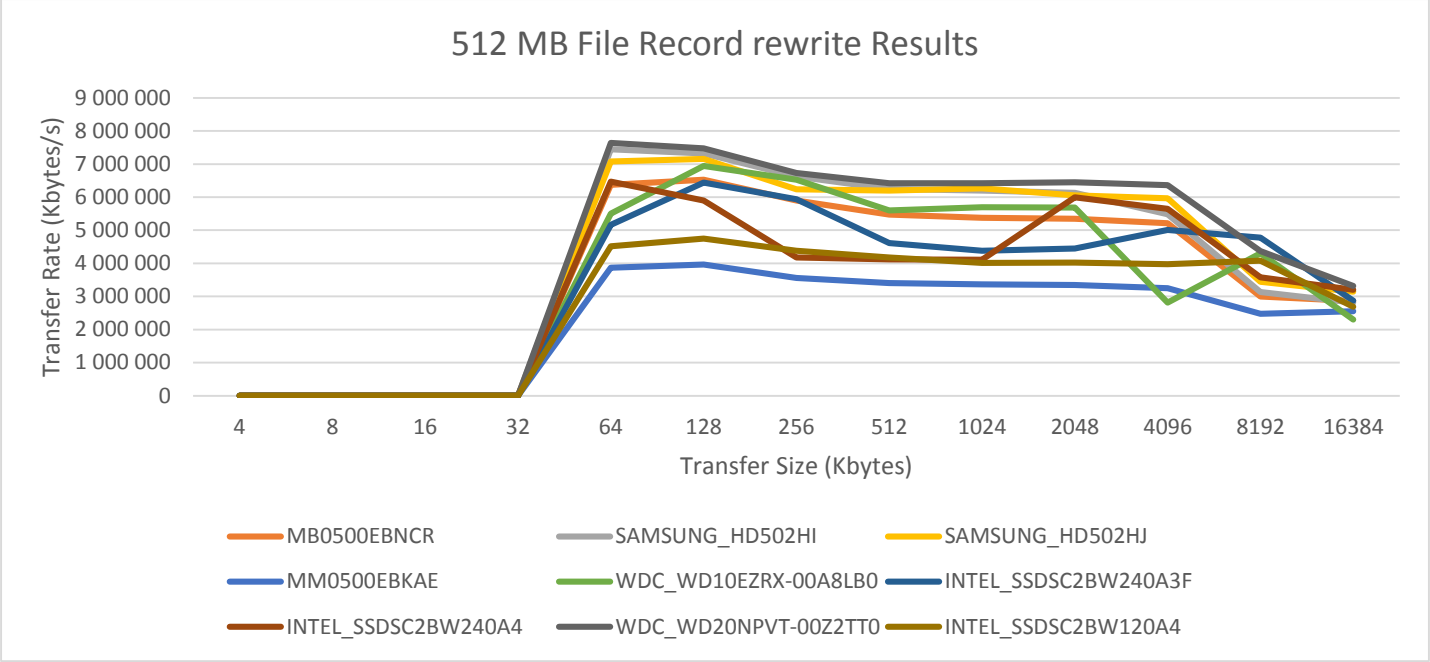
Record Rewrite

Este teste irá medir a performance na escrita e reescrita de um pedaço de informação dentro de um ficheiro. Podem acontecer coisas interessantes, se o pedaço for suficientemente pequeno que caiba na cache a performance será muito alta, sempre que aumentamos o tamanho do ficheiro a performance tenderá a diminuir pois será maior que as diferentes caches no sistema.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







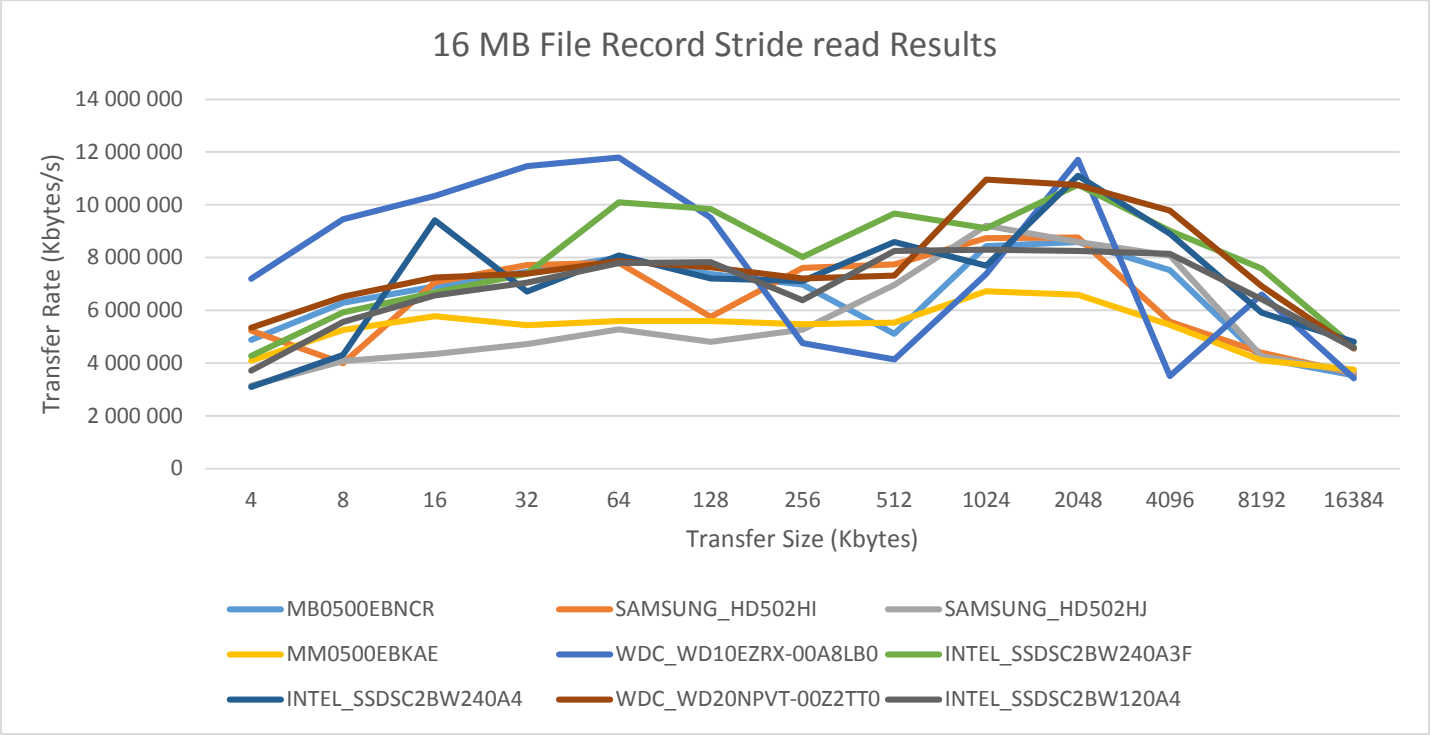
Conclusão

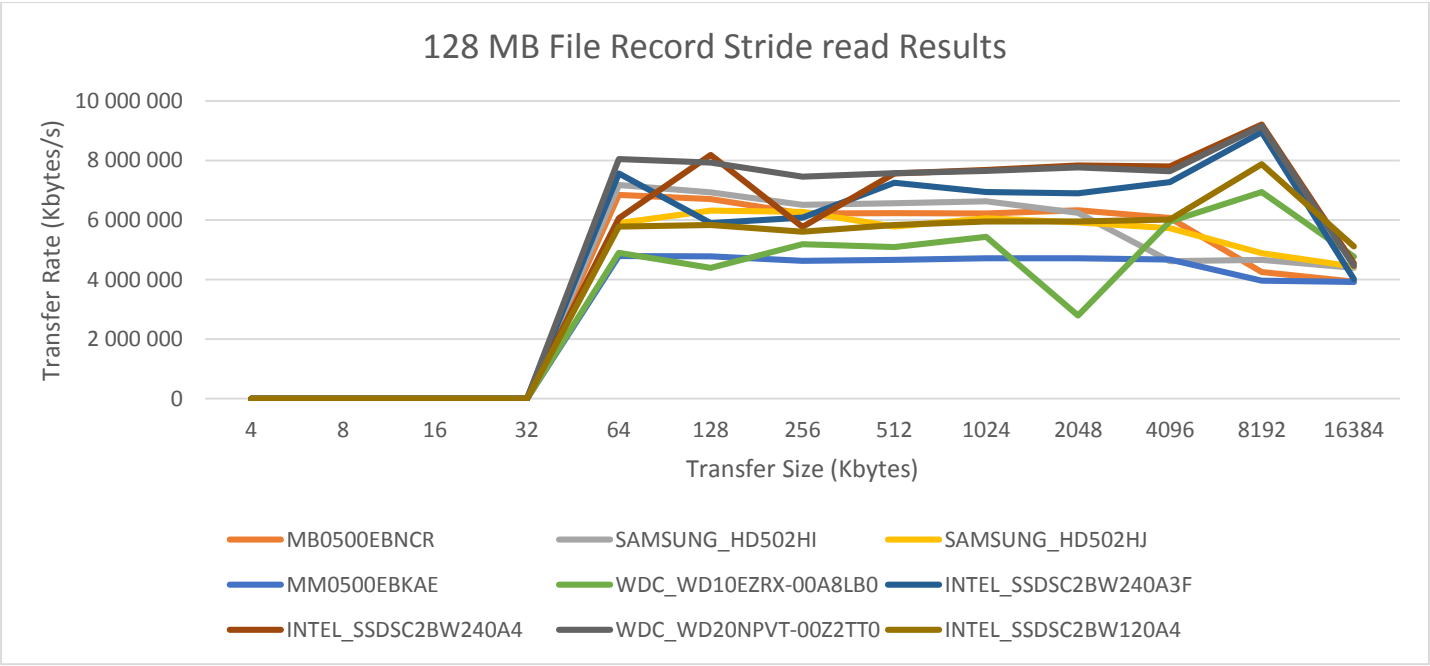
Por mais uma vez o **WDC_WD20NPVT-00Z2TT0** foi o que apresentou melhores resultados nos 3 tamanhos, resultados esses bastante constantes.

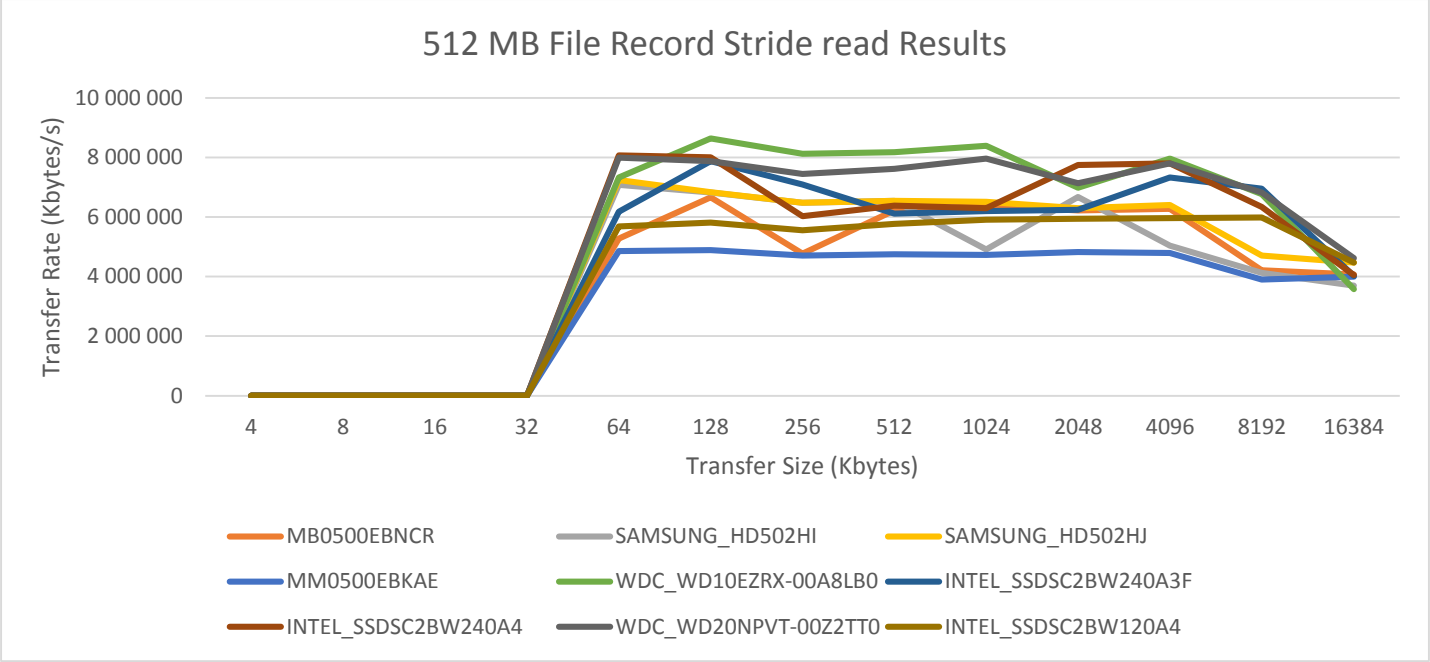
Strided Read

Este teste analisa a performance na leitura de um ficheiro com saltos de x Bytes como um padrão. Este acesso é usual em programas que acedem a regiões particulares em estruturas de dados. É uma acesso difícil de detetar pelo SO, produzindo interessantes anomalias de performance.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







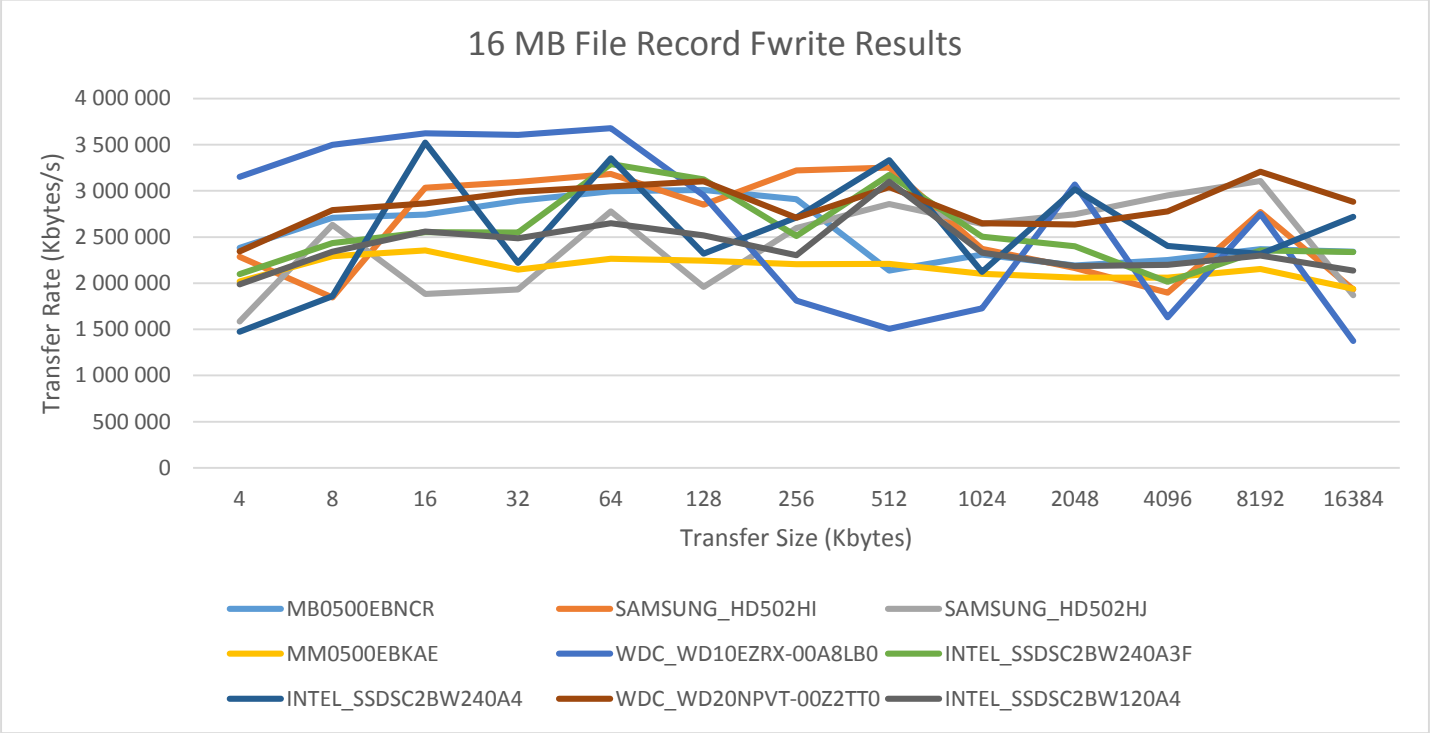
Conclusão

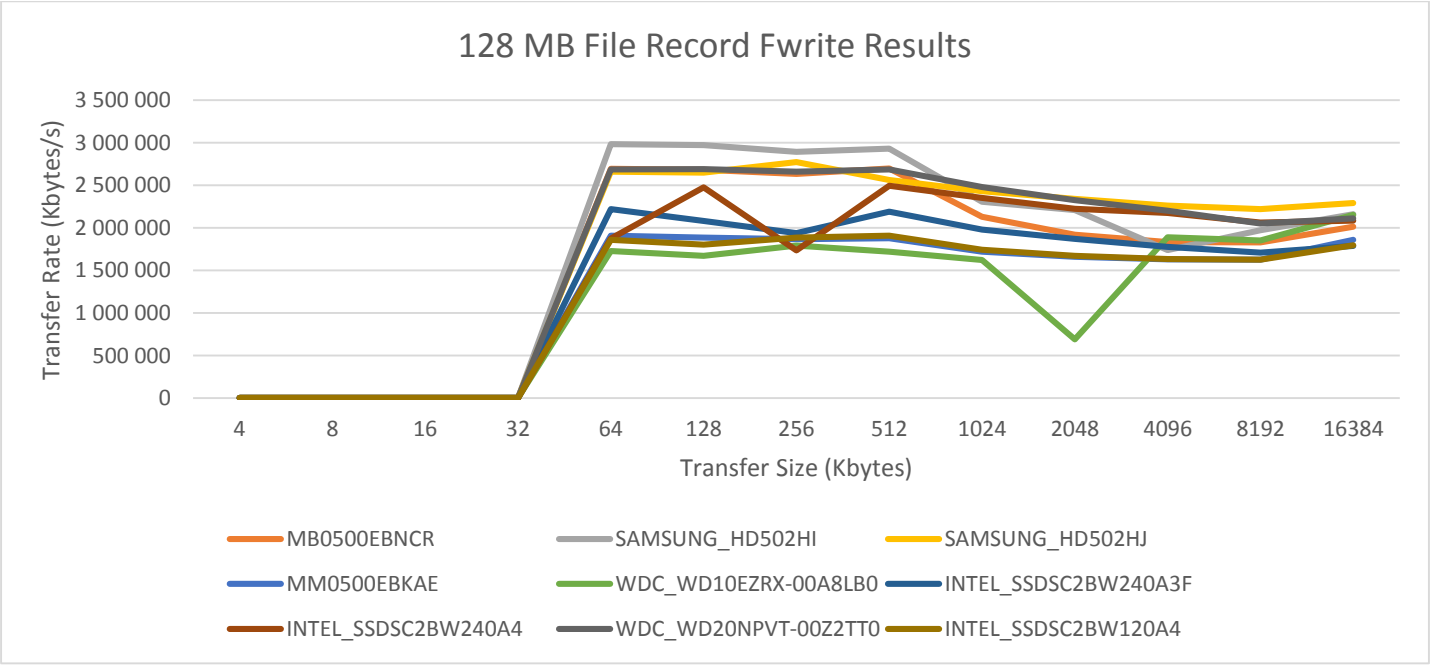
A oscilação de valores é muito acentuada neste teste.
O disco **WDC_WD10EZRX-00A8LB0** na totalidade obteve maiores resultados em 16MB e 512MB.
Para 128 MB o **WDC_WD20NPVT-00Z2TT0** foi o melhor.

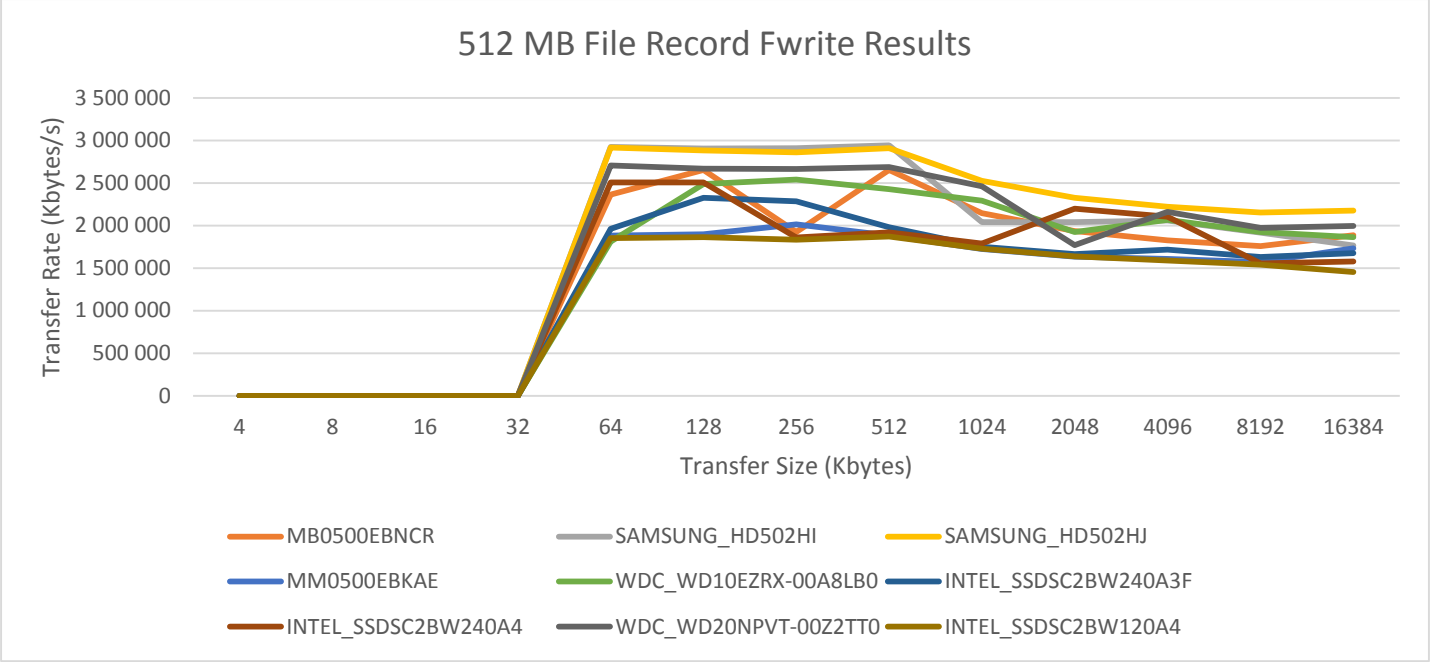
Fwrite

Este teste utiliza como operação principal a `fwrite()`. É uma rotina do sistema que permite fazer escritas com buffers, estando esse buffer em endereçamento do utilizador. Se a aplicação for escrever com transferências mais pequenas então as funcionalidades de buffer e bloqueio de E/S do *fwrite* podem aumentar a performance da aplicação reduzindo o número de chamadas ao sistema e aumentando o tamanho das transferências quando essas chamadas são feitas. O teste inclui a escrita de um ficheiro novo por isso os metadados são contabilizados.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







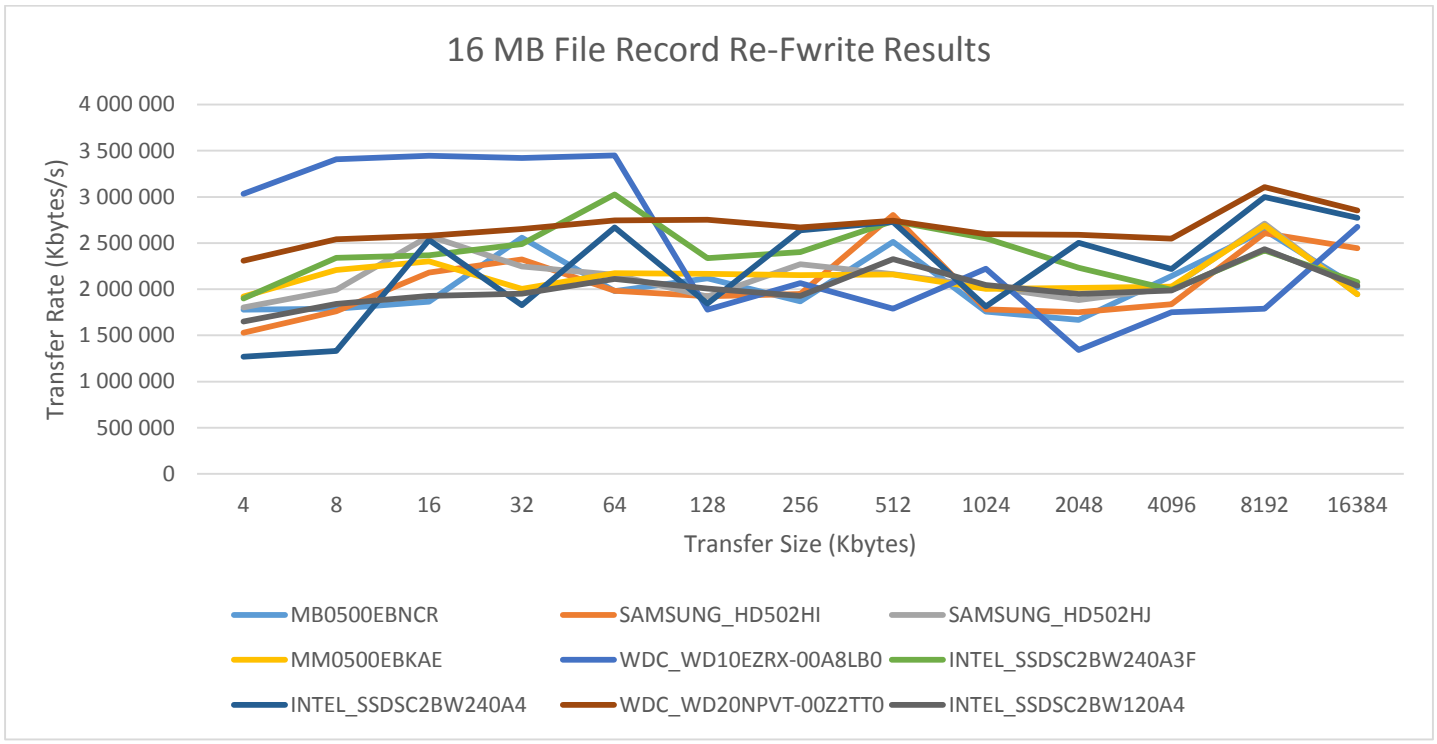
Conclusão

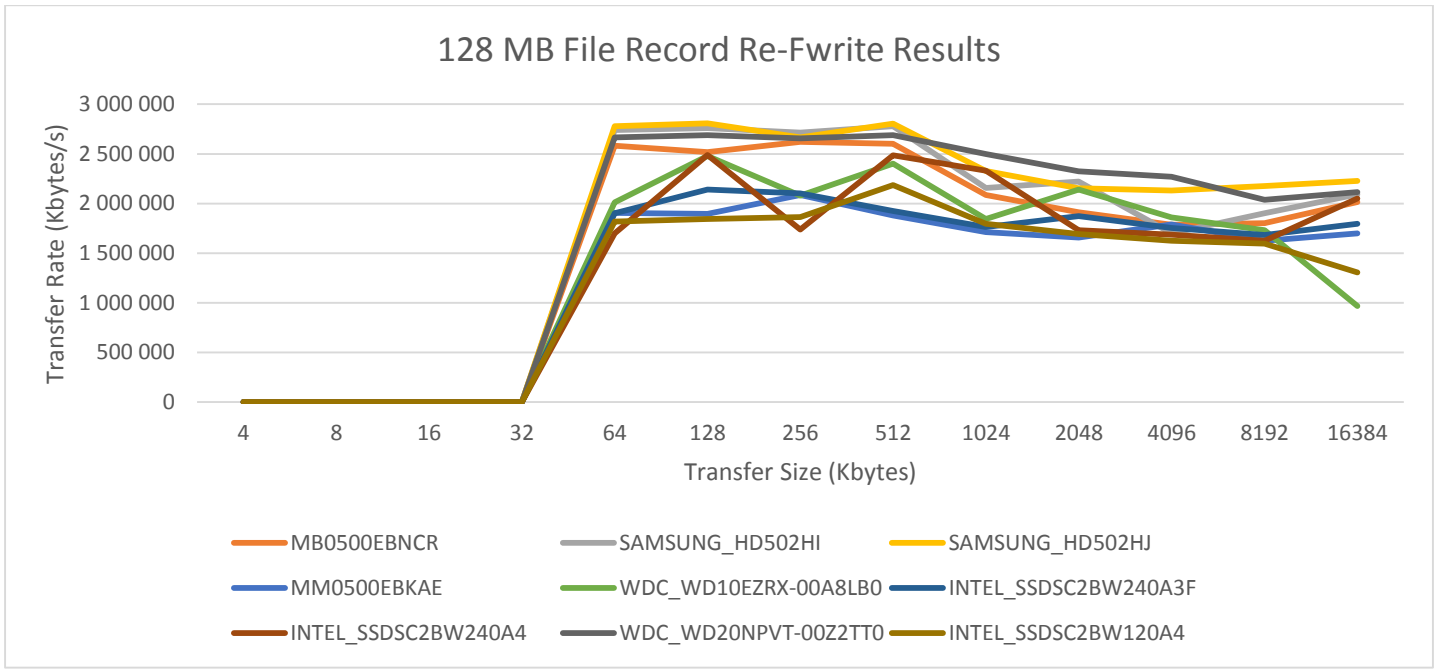
O disco **WDC_WD20NPVT-00Z2TT0** na totalidade obteve maiores resultados em 16MB.
Para os restantes tamanhos de teste o **SAMSUNG_HD502HJ** aparece como pioneiro em resultados.

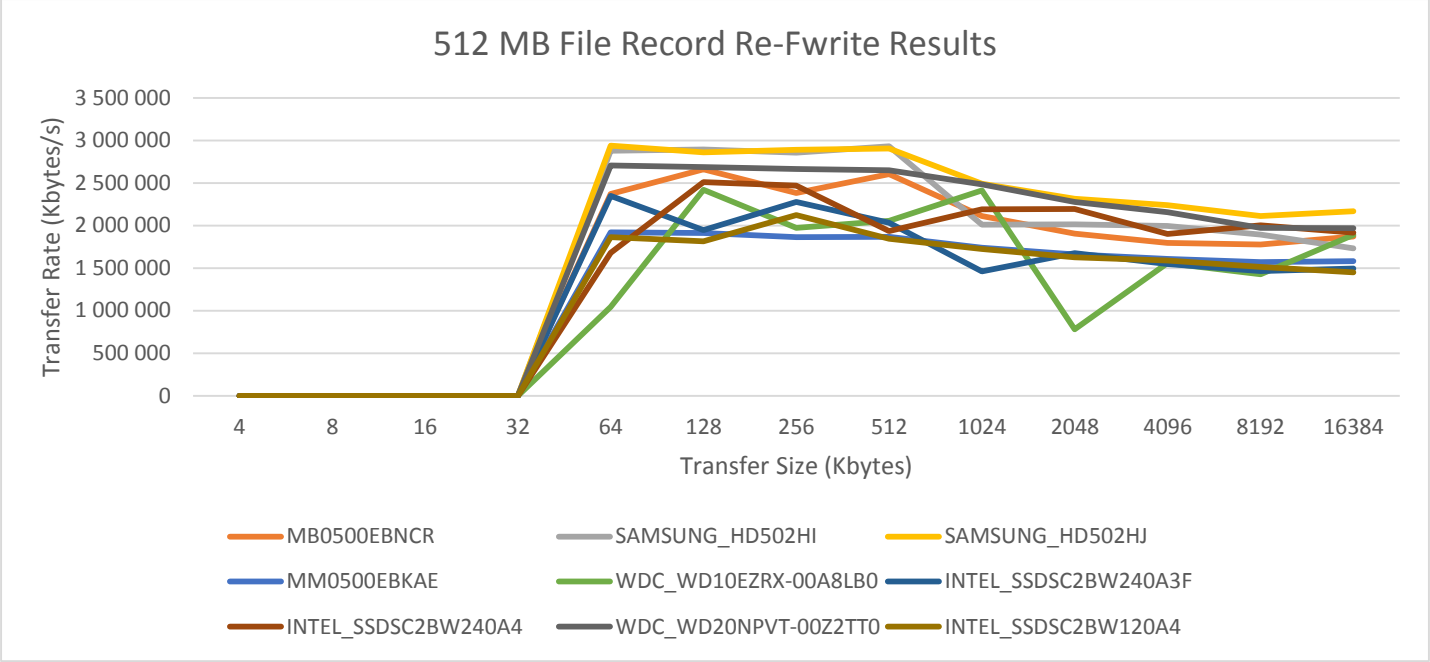
Re-Fwrite

Como o teste anterior, o *Re-Fwrite* utiliza o *fwrite* só que desta vez é escrever um ficheiro já existente por isso não há metadados envolvidos proporcionando teoricamente melhores resultados.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







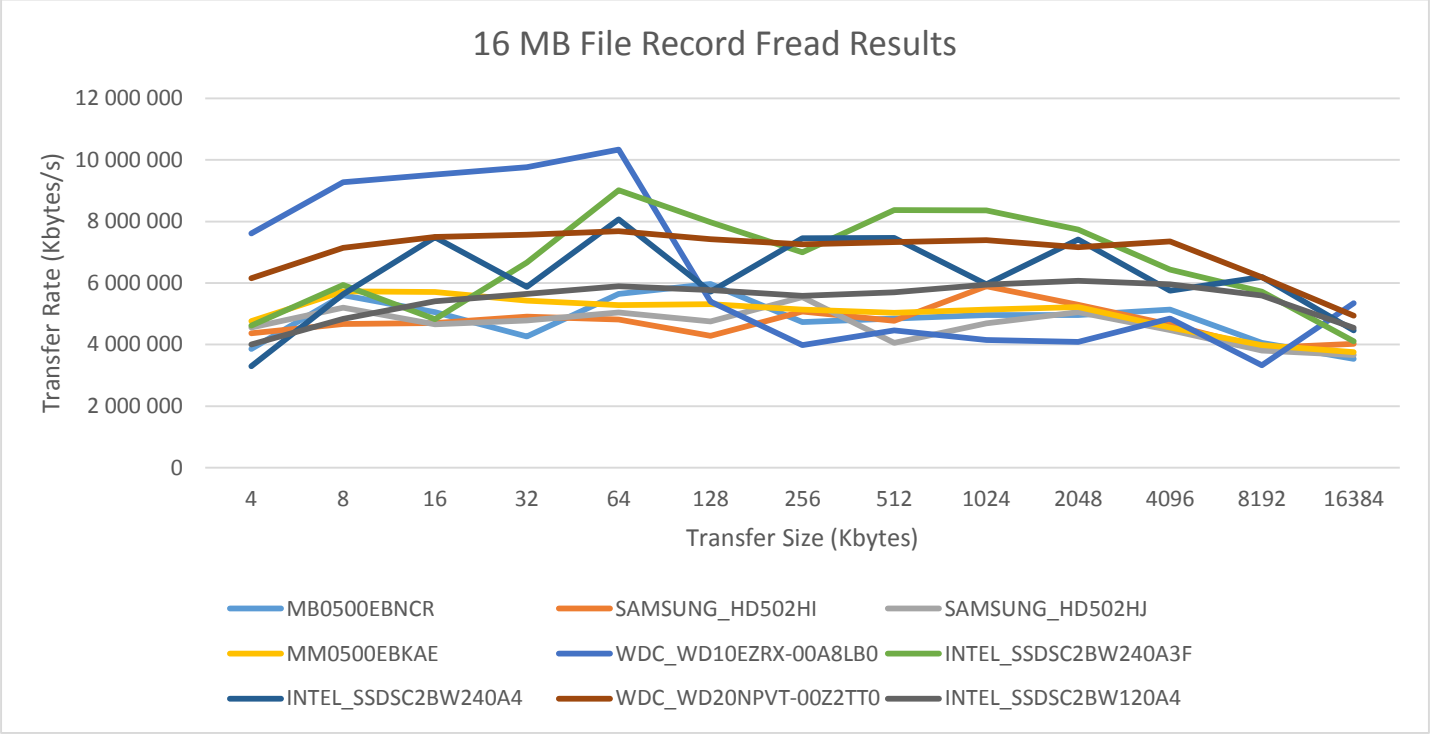
Conclusão

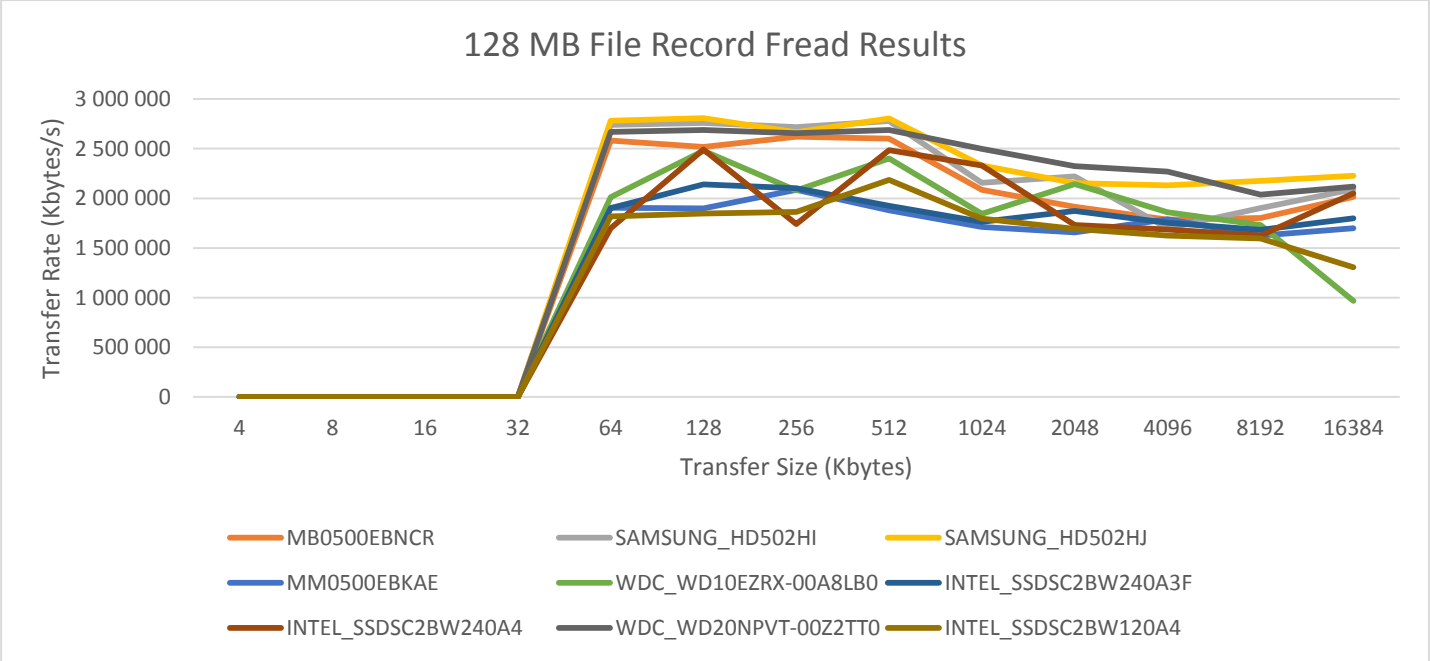
O disco **WDC_WD20NPVT-00Z2TT0** na totalidade obteve melhores resultados em 16MB e 128 MB.

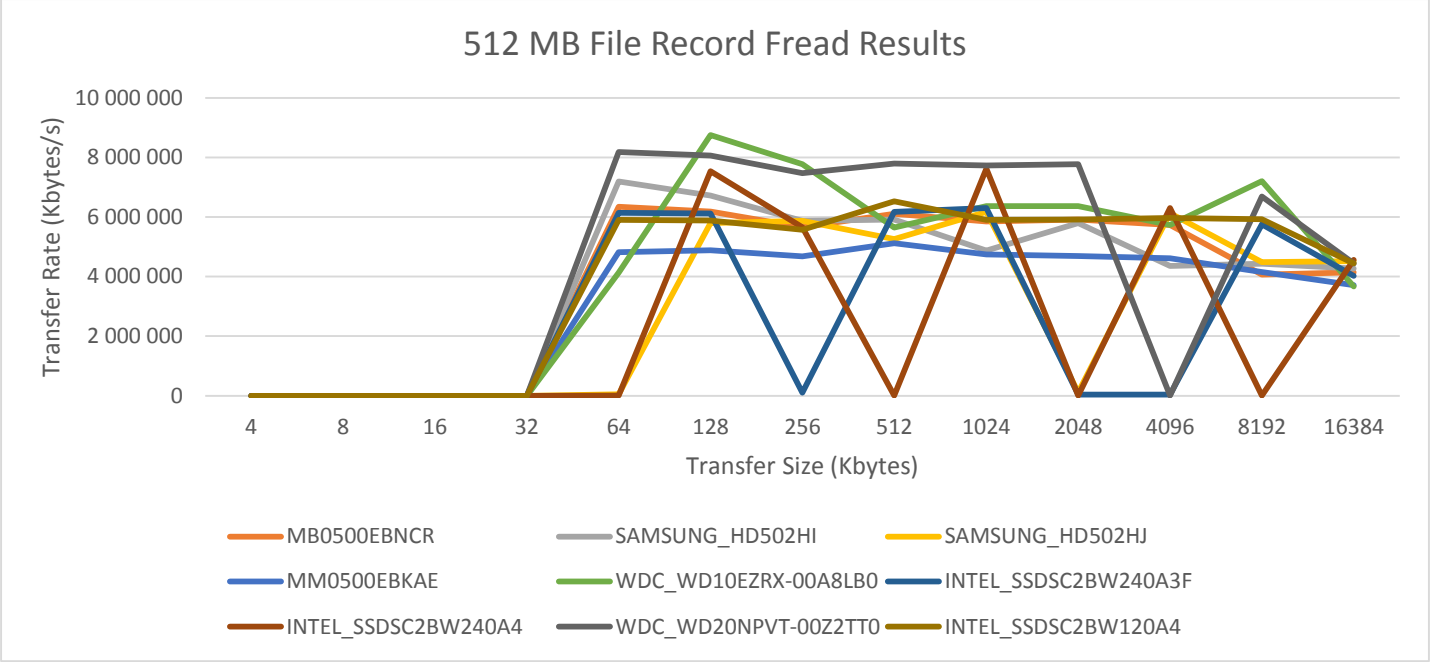
Já para 512MB o **SAMSUNG_HD502HJ** foi ligeiramente melhor que o vencedor dos 2 tamanhos anteriores.

Fread

Este teste utiliza a função `fread` que utiliza operações com buffer e bloqueios.
Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







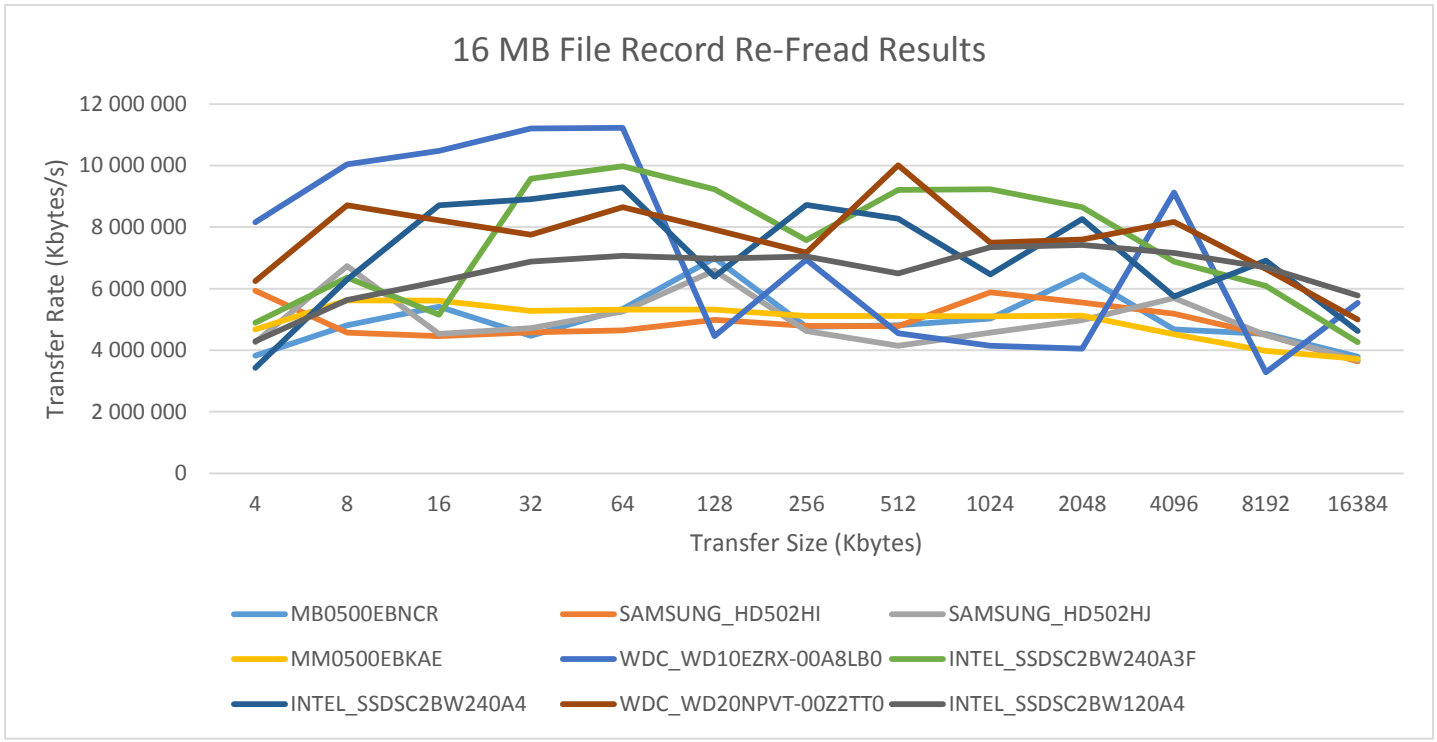
Conclusão

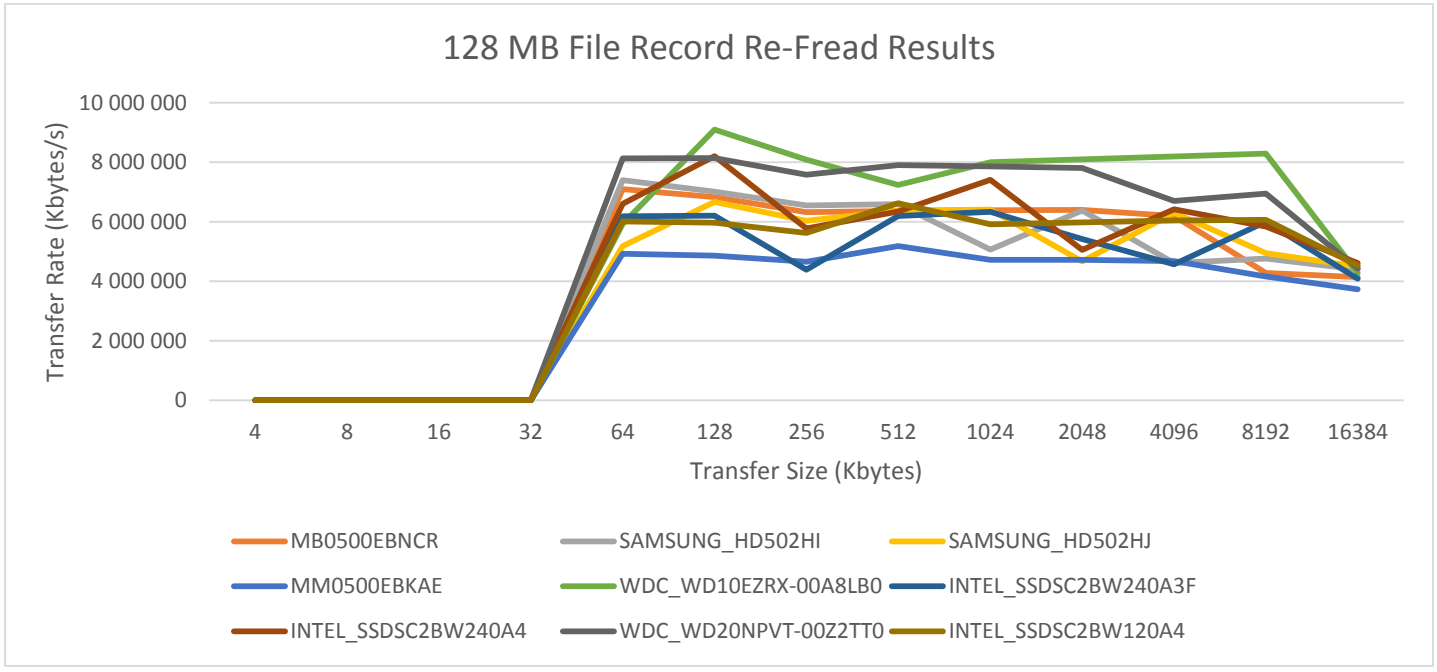
O disco **WDC_WD20NPVT-00Z2TT0** analiticamente foi superior em 16MB e 128 MB, o **SAMSUNG_HD502HJ** esteve com resultados muito próximos mas só foi efetivamente superior ao WDC para 512 MB.

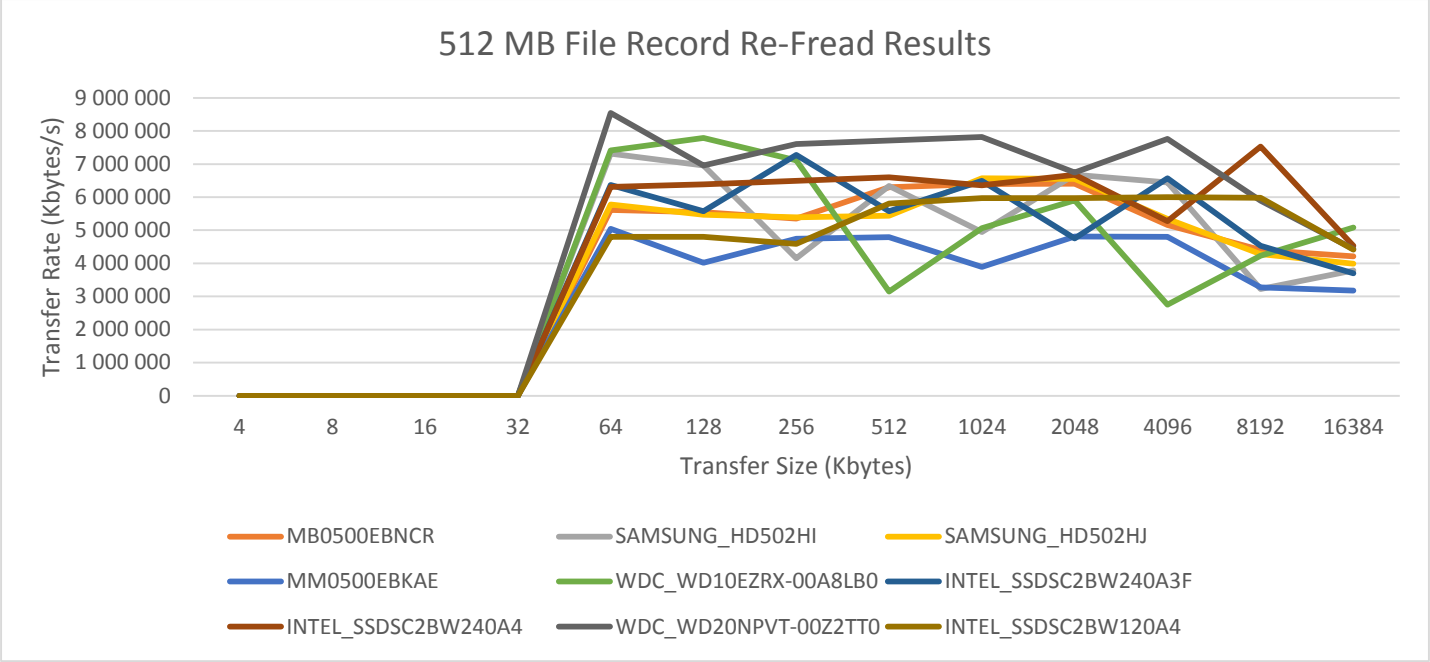
Re-Fread

Tal como o teste anterior a operação `fread` é utlizada, mas agora é lendo um ficheiro já lido previamente esperando que já estejam dados do ficheiro em cache.

Com os valores obtidos retirei 3 gráficos para 16 MB, 128 MB e 512 MB.







Conclusão

O disco **WDC_WD20NPVT-00Z2TT0** foi superior em 16MB e 128 MB, já o **WDC_WD10EZRX-00A8LB0** apenas o superou nos 512 MB.

Scripts, Diretorias e Gráficos

Cada nó tem uma pasta respetiva, contendo os ficheiros output do IOzone em .txt, ficheiros excel em .xls e ainda os gráficos gerados separados por pasta 0 (até 512MB) e 8 (até 8GB).

|-- 431-3 | |-- 0 | `-- 8 |-- 431-5 |  |-- 0 |  `-- 8 |-- 431-6 | |-- 0 | `-- 8	|-- 432-1 | |-- 0 | `-- 8 |-- 541-1 |  |-- 0 |  `-- 8 |-- 641-19 | |-- 0 | `-- 8	|-- 641-8 | |-- 0 | `-- 8 |-- 652-1 |  |-- 0 |  `-- 8 |-- 662-6 | |-- 0 | `-- 8

Os Gráficos foram gerados graças à ferramenta disponibilizada intitulada de *Generate_Graphs* que utiliza o *gnuplot*. Os gráficos individuais não foram inseridos no relatório pois são muitos (117 gráficos só para o teste de 512 MB !), portanto estão nas pastas para consulta posterior.

Os testes de 8GB não terminaram como previsto para todos os nós portanto não pude utilizar esses valores para análise.

Análise Final de Resultados e Conclusão

O teste do IOzone é muito intensivo operacionalmente e demorou várias horas para concluir os testes. Após foi necessário analisar os valores e gerar gráficos comparativos para todos os discos analisados.

O disco que obteve melhores resultados foi claramente o **WDC_WD20NPVT-00Z2TT0** pertencente aos nós 652-1 e 652-2. Estava à espera que um SSD tivesse os resultados com clara liderança mas tal não aconteceu, apenas obtendo 2 testes vitoriosos com SSD’s diferentes.