Attribution 4.0 International (CC BY 4.0)

https://creativecommons.org/licenses/by/4.0/ (https://creativecommons.org/licenses/by/4.0/)

https://en.wikipedia.org/wiki/S.M.A.R.T.#Known_ATA_S.M.A.R.T._attributes (https://en.wikipedia.org/wiki/S.M.A.R.T.#Known_ATA_S.M.A.R.T._attributes)

In [1]:
```
import pandas
import sys
print(sys.version_info)
print('pandas',pandas.__version__)
import glob
import pickle
import numpy
import time
import matplotlib.pyplot as plt
```

```
sys.version_info(major=3, minor=6, micro=6, releaselevel='final', seria
l=0)
pandas 0.23.4
```

In [2]:
```
df_header_only=pandas.read_csv('zipped_data/data_Q2_2018.zip_folder/2018
-04-01.csv',nrows=3)
nonsmart_cols=[]
for colname in df_header_only.columns:
    if 'smart_' not in colname:
        nonsmart_cols.append(colname)
```

In [3]:
```
nonsmart_cols.append('smart_241_raw') # written
nonsmart_cols.append('smart_242_raw') # read
nonsmart_cols.append('smart_9_raw') # power-on hours
nonsmart_cols.remove('capacity_bytes')
```

In [4]:
```
list_of_csvs = glob.glob('zipped_data/**/*.csv', recursive=True)
len(list_of_csvs)
```

Out[4]:
```
2092
```

In [5]:
```
start_time=time.time()
list_of_df=[]
for csv_file in list_of_csvs:
    df=pandas.read_csv(csv_file,nrows=2)
    if 'smart_241_raw' in df.columns:
        df=pandas.read_csv(csv_file,usecols=nonsmart_cols)
        df = df[df['failure']==1]
        list_of_df.append(df)
print('elapsed:',time.time()-start_time,'seconds')
```

```
elapsed: 446.1708550453186 seconds
```

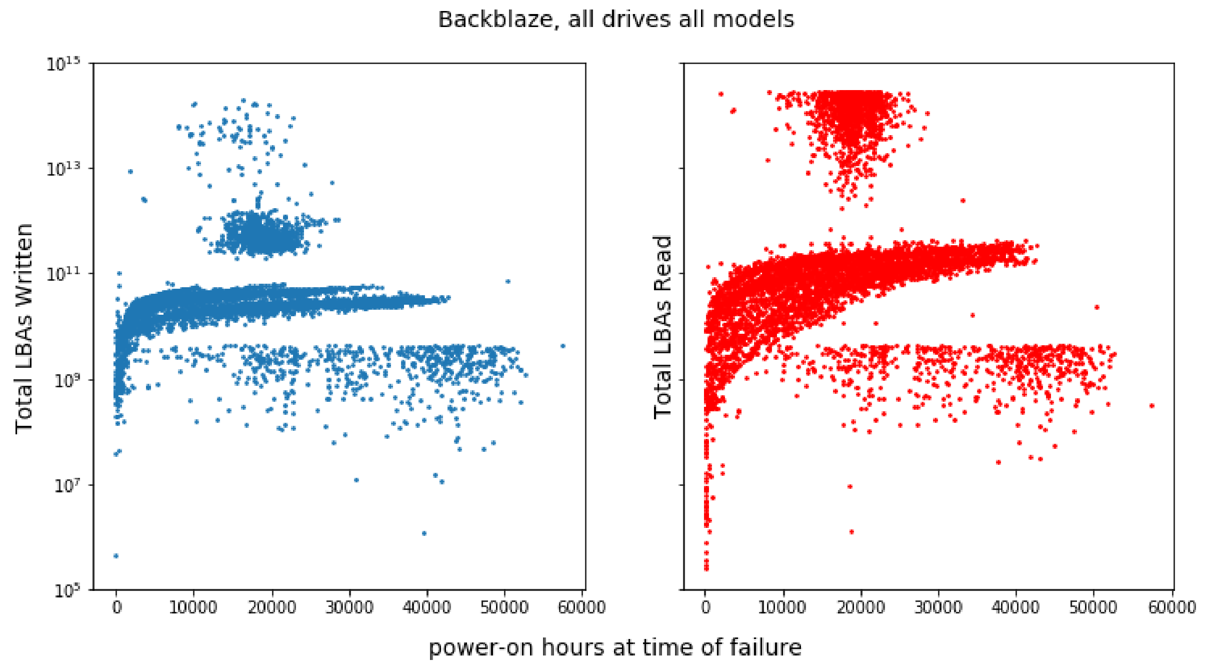
In [6]:
```
df = pandas.concat(list_of_df)
print(df.shape)
#df.dropna(how='any',inplace=True)
#print(df.shape)
#df.head()
```

```
(8743, 7)
```

Logical Block Addresses read/written versus power-on hours

In [59]:
```
def lba_plot_vs_poh(y_lower,y_upper,log_bool):
    f, (ax1, ax2) = plt.subplots(1, 2, sharey=True,figsize=(12, 6))
    ax1.scatter(x=df['smart_9_raw'],y=df['smart_241_raw'],s=3)
    ax1.set_ylabel('Total LBAs Written',fontsize=14)
    ax1.set_yscale('log')

    ax2.scatter(x=df['smart_9_raw'],y=df['smart_242_raw'],color='r',s=3)
    ax2.set_ylabel('Total LBAs Read',fontsize=14);
    plt.ylim([y_lower,y_upper])
    if log_bool:
        ax2.set_yscale('log')

    f.text(0.5, 0.04, 'power-on hours at time of failure', ha='center',
va='center',fontsize=14);
    f.text(0.5, 0.94, 'Backblaze, all drives all models', ha='center', v
a='center',fontsize=14);
    return
```

In [60]:
```
lba_plot_vs_poh(y_lower=100000,y_upper=1000000000000000,log_bool=True)
```



zoom in to the "low LBA read/written" range of values

lba_plot_vs_poh(y_lower=0,y_upper=10000000000,log_bool=True)

zoom out to the "medium LBA read/written" range of values

lba_plot_vs_poh(y_lower=0,y_upper=100000000000,log_bool=False)

zoom out again to the "high LBA read/written" range of values

lba_plot_vs_poh(y_lower=0,y_upper=1000000000000,log_bool=False)

max range for y-axis

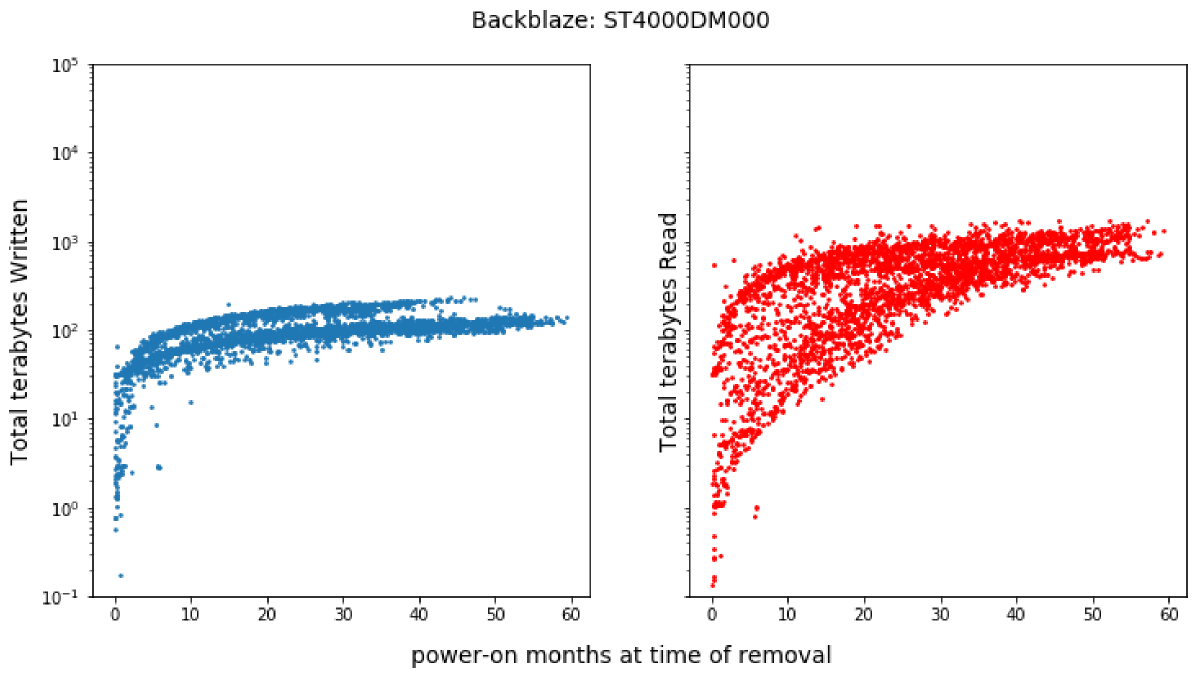
lba_plot_vs_poh(y_lower=0,y_upper=1E15,log_bool=True)

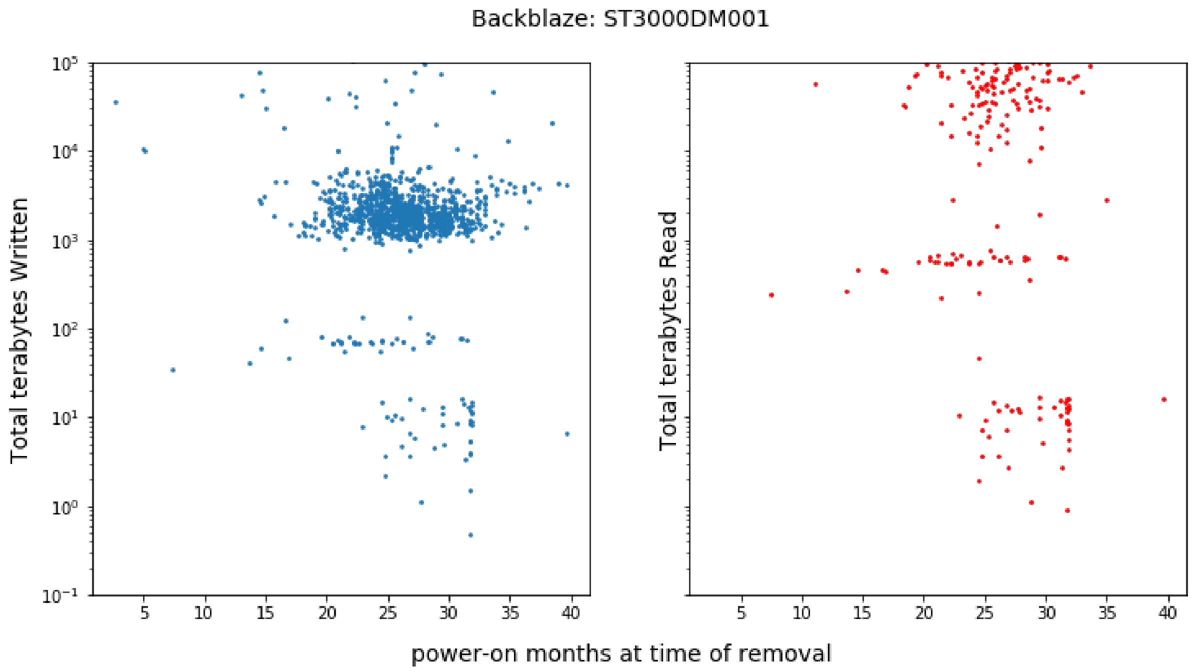
per model

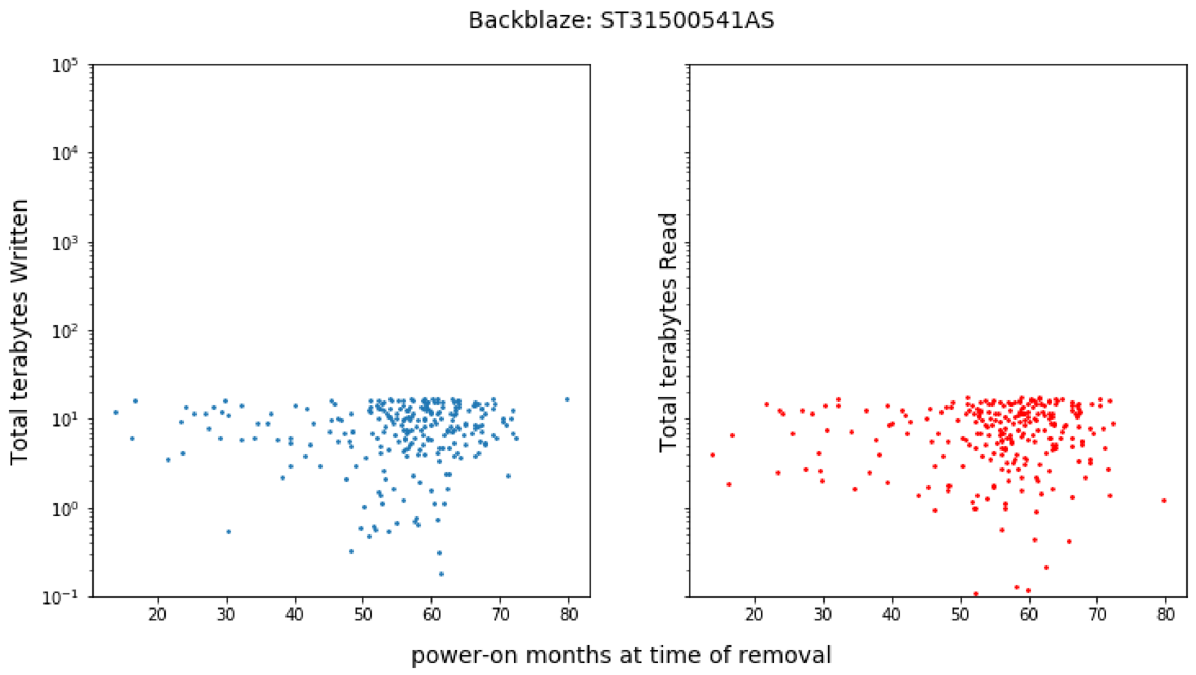
def make_plot_per_model(drive_model): if not (df[df['model']==drive_model]['smart_241_raw'].isnull().all()): f, (ax1, ax2) = plt.subplots(1, 2, sharey=True,figsize=(12, 6)) ax1.scatter(x=df[df['model']==drive_model]['smart_9_raw'],y=df[df['model']==drive_model]['smart_241_raw'], s=3,label=drive_model) ax1.set_ylabel('Total Logical Block Addresses Written',fontsize=14) ax1.set_yscale('log') ax2.scatter(x=df[df['model']==drive_model]['smart_9_raw'],y=df[df['model']==drive_model]['smart_242_raw'],color='r', s=3,label=drive_model) ax2.set_ylabel('Total Logical Block Addresses Read',fontsize=14); plt.ylim([100000,1000000000000000]) ax2.set_yscale('log') f.text(0.5, 0.04, 'power-on hours at time of removal', ha='center', va='center',fontsize=14); f.text(0.5, 0.94, 'Backblaze: '+drive_model, ha='center', va='center',fontsize=14);drive_model='ST4000DM000' print(df[df['model']==drive_model].shape) make_plot_per_model(drive_model)ser = df['model'].value_counts() for drive_model in ser[ser>100].index: # only show results if there are more than 100 instances of that drive model being removed make_plot_per_model(drive_model)

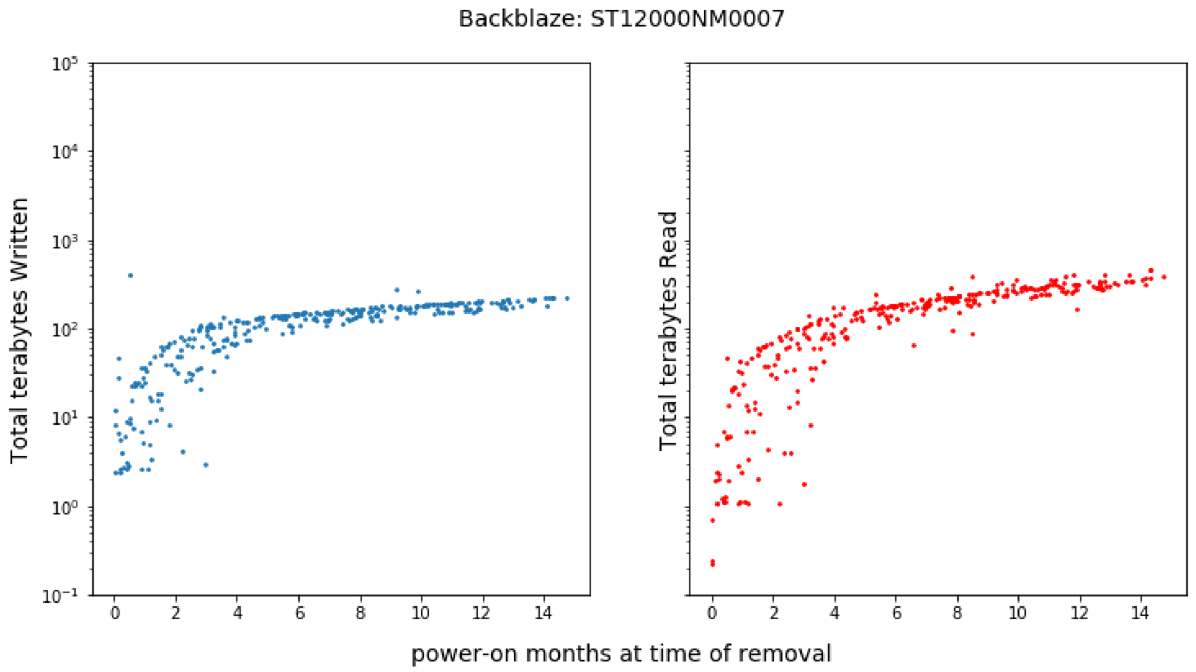
In [61]:
```
# LBA*(4*1012) bytes to tb

def make_TB_plot_per_model(drive_model):

    if not (df[df['model']==drive_model]['smart_241_raw'].isnull().all
()):
        f, (ax1, ax2) = plt.subplots(1, 2, sharey=True,figsize=(12, 6))
        ax1.scatter(x=df[df['model']==drive_model]['smart_9_raw']/24/30,
y=df[df['model']==drive_model]['smart_241_raw']*(4*1012)/(1E12),
                    s=3,label=drive_model)
        ax1.set_ylabel('Total terabytes Written',fontsize=14)
        ax1.set_yscale('log')

        ax2.scatter(x=df[df['model']==drive_model]['smart_9_raw']/24/30,
y=df[df['model']==drive_model]['smart_242_raw']*(4*1012)/(1E12),color=
'r',
                    s=3,label=drive_model)
        ax2.set_ylabel('Total terabytes Read',fontsize=14);
        plt.ylim([0.1,100000])
        ax2.set_yscale('log')

        f.text(0.5, 0.04, 'power-on months at time of removal', ha='cent
er', va='center',fontsize=14);
        f.text(0.5, 0.94, 'Backblaze: '+drive_model, ha='center', va='ce
nter',fontsize=14);
```

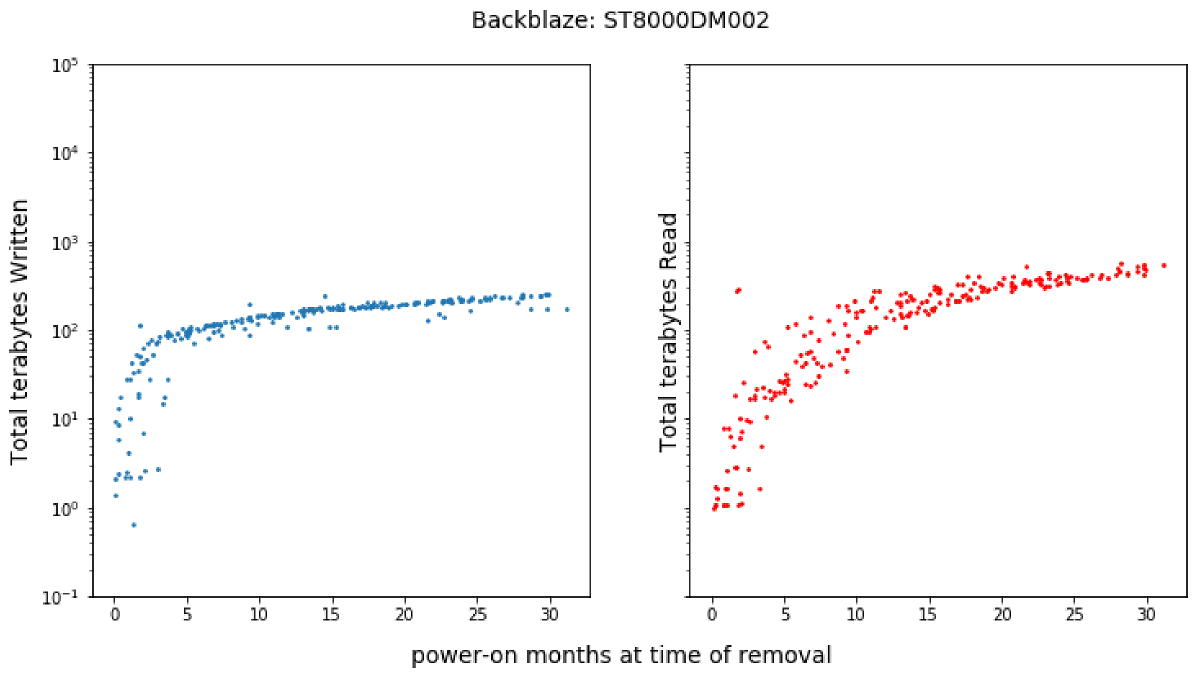
In [62]:
```
ser = df['model'].value_counts()
for drive_model in ser[ser>100].index: # only show results if there are
 more than 100 instances of that drive model being removed
    make_TB_plot_per_model(drive_model)
```

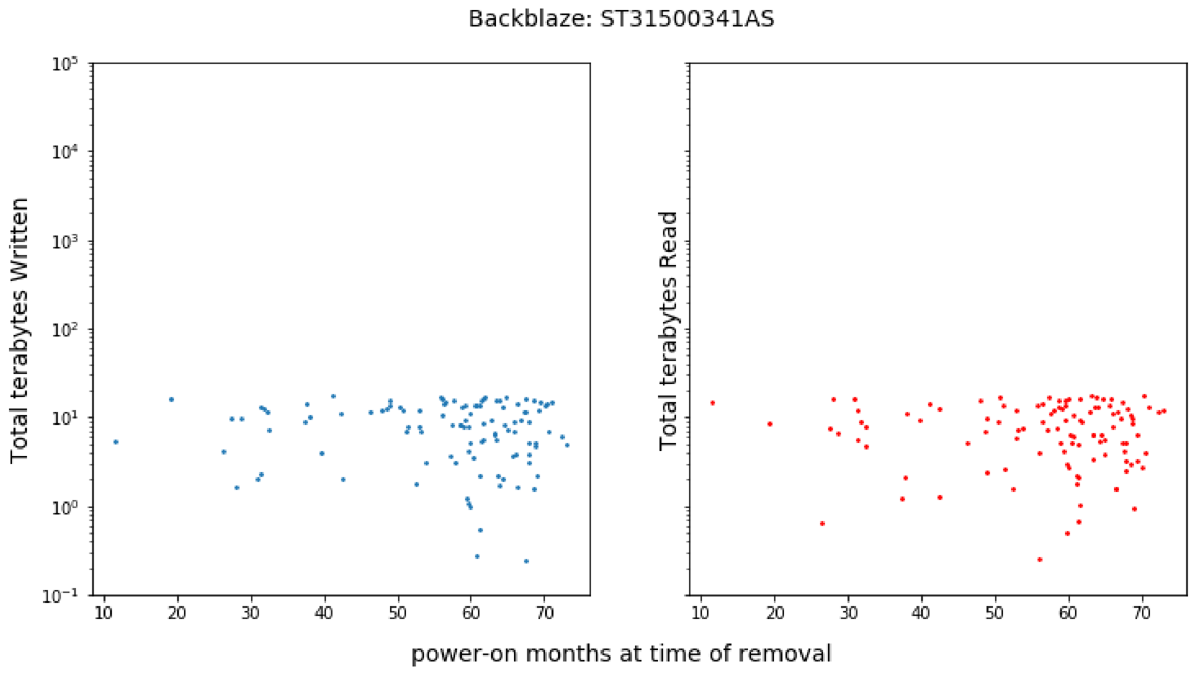

Backblaze: ST4000DM000
Total terabytes Written
Total terabytes Read
power-on months at time of removal
10^5
10^4
10^3
10^2
10^1
10^0
10^-1
0
10
20
30
40
50
60

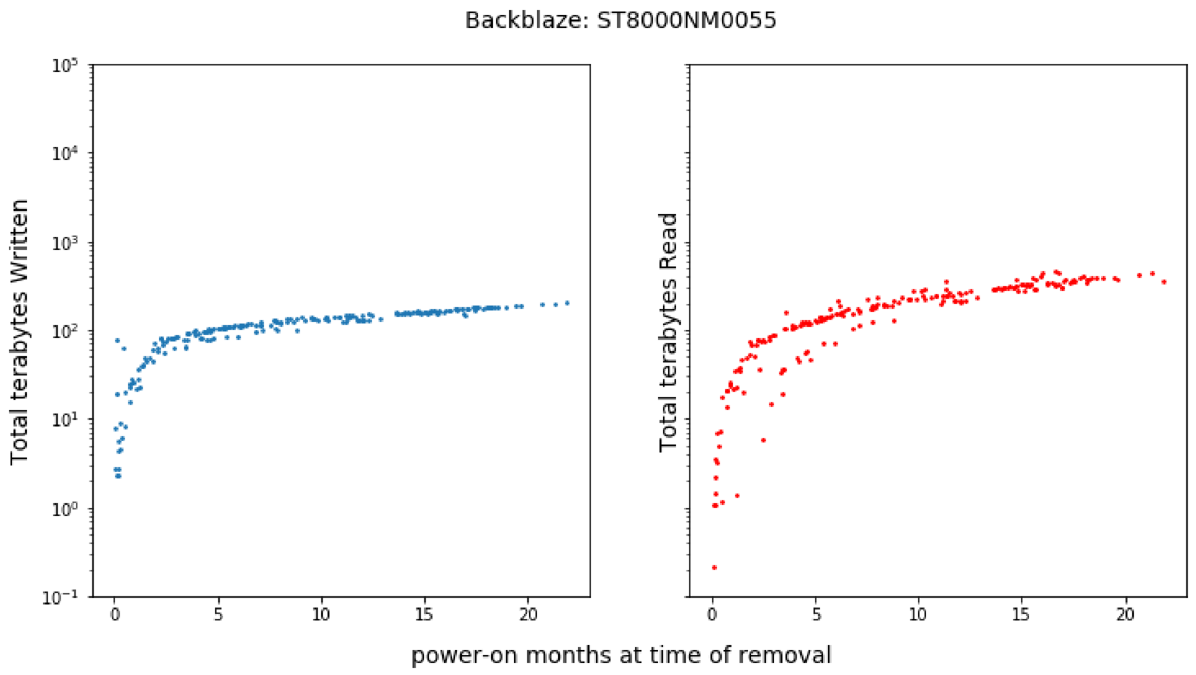

Backblaze: ST3000DM001
Total terabytes Written
Total terabytes Read
power-on months at time of removal
10^5
10^4
10^3
10^2
10^1
10^0
10^-1
5
10
15
20
25
30
35
40


Backblaze: ST31500541AS
Total terabytes Written
Total terabytes Read
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
20
30
40
50
60
70
80
power-on months at time of removal


Backblaze: ST12000NM0007
Total terabytes Written
Total terabytes Read
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
0
2
4
6
8
10
12
14
power-on months at time of removal


Backblaze: ST8000DM002
Total terabytes Written
Total terabytes Read
power-on months at time of removal
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
0
5
10
15
20
25
30


Backblaze: ST31500341AS
Total terabytes Written
Total terabytes Read
power-on months at time of removal
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10
20
30
40
50
60
70


Backblaze: ST8000NM0055
Total terabytes Written
Total terabytes Read
power-on months at time of removal
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
0
5
10
15
20