# RateAnalysis

October 7, 2019

## 1 Introduction

Validating EBDC online compression throughput using the Supermicro SuperWorkstation 7049GP-TRT, with 2 x Intel Xeon Silver 4216 Processor 16-Core 2.1GHz 32 core CPUs and 128 GB memory.

The data is all 2019 sPHENIX TPC SAMPA data at FTBF total 1+TB. The data are buffered on ASUS Hyper M.2 X16 PCIe 3.0 X4 Expansion Card V2 with four SAMSUNG 970 EVO PLUS M.2 2280 1TB PCIe Gen 3.0 x4 NVMe 1.3 V-NAND configured in 4-strip software RAID0. The RAID is tested to 6GBps write and 11GBps write through its PCIe Gen3 x16 interface, matching a large fraction of the FELIX throughput and suppass the expected average rate in sPHENIX year-5 operation.

The data is readout as `parallel` jobs via start-compression.sh, and sink via either `/dev/null` or TPC connections to multiple `ncat` processes either at `localhost` or remote which can be started with start-sink.sh

## 2 Inputs

[11]:
```
# DataDir = './data_tmp/'
# DataDir = './data_25x_localhost/'
# DataDir = './data_48x_localhost/'
# studytitle = r"$\bf{EBDC}$" + " compression\nlocalhost loopback"
# DataDir = './data_48x_null/'
# studytitle = r"$\bf{EBDC}$" + " compression\noutput to /dev/null"



DataDir = './data_48x_20GbpsNetwork/'
studytitle = r"$\bf{EBDC}$" + " compression\n2x10Gbps ethernet"
# DataDir = './data_25x_20GbpsNetwork/'
# studytitle = r"$\bf{EBDC}$" + " compression\n2x10Gbps ethernet"
```

[12]:
```
# %matplotlib widget
# %matplotlib ipympl
%matplotlib inline
# well the html export like dump formats
```

## 3 Processing

[13]:
```
import os
import ntpath
import re
import pandas as pd
import numpy as np


def processDataset(dataset: str):
    split = dataset.split('-')
    if (len(split) != 3):
        print('skip {}'.format (dataset) );
        return;

    zipcmd = split[0];
    ziplevel = int(split[1]);
    jobs = int(split[2]);
    datasetDir = os.path.join(DataDir, dataset)

    print('processing {}, {} level{} x{}'.format (datasetDir,␣
 ↪zipcmd,ziplevel,jobs) );

    datasubfolders = [os.path.basename(f.path) for f in os.scandir(datasetDir)␣
 ↪if f.is_file() ]
    datasubfolders.sort()
    rpv_in = re.compile('pv_in_([0-9]*)\.log')

    for data in datasubfolders:
#         print ('data = {}'.format(data));
        m = rpv_in.search(data)
        if m is not None:
#             print ('found {} -> {}'.format(data, m.group(1)));
            jobID = m.group(1);
            with open(os.path.join(datasetDir, 'pv_in_{}.log'.format(jobID)))␣
 ↪as f:
                split = f.readlines()[-1].split();
                assert(len(split)==2)
                inTime = float(split[0])
                inSize = float(split[1])
            with open(os.path.join(datasetDir, 'pv_out_{}.log'.format(jobID)))␣
 ↪as f:
                split = f.readlines()[-1].split();
                assert(len(split)==2)
                outTime = float(split[0])
                outSize = float(split[1])
```

```
#              print ('df.append {} . {} , {} -> {}'.format(data,␣
↪jobID,inSize,outSize));
            dictData = { 'dataset' : dataset ,
                       'zipcmd': zipcmd,
                       'ziplevel': ziplevel,
                       'jobs': jobs,
                       'jobID': jobID,
                       'inTime': inTime,
                       'inSize': inSize,
                       'outTime': outTime  ,
                       'outSize': outSize
                       }
            global dataframe
            dataframe = dataframe.append(dictData, ignore_index=True)



dataframe = pd.DataFrame(columns=['dataset', 'zipcmd', 'ziplevel', 'jobs',␣
↪'jobID', 'inTime', 'inSize', 'outTime', 'outSize'])
subfolders = [f.path for f in os.scandir(DataDir) if f.is_dir() ]
subfolders.sort()

for dataset in subfolders:
    processDataset(os.path.basename(dataset))
```

```
processing ./data_25x_20GbpsNetwork/gzip-1-25, gzip level1 x25
processing ./data_25x_20GbpsNetwork/gzip-2-25, gzip level2 x25
processing ./data_25x_20GbpsNetwork/gzip-3-25, gzip level3 x25
processing ./data_25x_20GbpsNetwork/gzip-5-25, gzip level5 x25
processing ./data_25x_20GbpsNetwork/gzip-7-25, gzip level7 x25
processing ./data_25x_20GbpsNetwork/gzip-9-25, gzip level9 x25
processing ./data_25x_20GbpsNetwork/lz4-1-25, lz4 level1 x25
processing ./data_25x_20GbpsNetwork/lz4-2-25, lz4 level2 x25
processing ./data_25x_20GbpsNetwork/lz4-3-25, lz4 level3 x25
processing ./data_25x_20GbpsNetwork/lz4-5-25, lz4 level5 x25
processing ./data_25x_20GbpsNetwork/lz4-7-25, lz4 level7 x25
processing ./data_25x_20GbpsNetwork/lz4-9-25, lz4 level9 x25
processing ./data_25x_20GbpsNetwork/lzop-1-25, lzop level1 x25
processing ./data_25x_20GbpsNetwork/lzop-2-25, lzop level2 x25
processing ./data_25x_20GbpsNetwork/lzop-3-25, lzop level3 x25
processing ./data_25x_20GbpsNetwork/lzop-5-25, lzop level5 x25
processing ./data_25x_20GbpsNetwork/lzop-7-25, lzop level7 x25
processing ./data_25x_20GbpsNetwork/lzop-9-25, lzop level9 x25
```

## 4 Plot

[14]:
```
dataframeSum = pd.DataFrame(columns=['dataset', 'zipcmd', 'ziplevel',␣
 ↪'jobs','totalInTime', 'totalInSize', 'totalOutTime', 'totalOutSize',␣
 ↪'Compression', 'inRateGbps', 'outRateGbps'])

zipcmds = dataframe.zipcmd.unique()

for zipcmd in zipcmds:

    zipRows = dataframe.loc[dataframe['zipcmd'] == zipcmd]

    ziplevels = zipRows.ziplevel.unique()
    for ziplevel in ziplevels:
        ziplevelRows = zipRows.loc[zipRows['ziplevel'] == ziplevel]
        print ('processing ', zipcmd, '.',ziplevel, ' size= ',ziplevelRows.
 ↪size, 'compression ratio = ',ziplevelRows['outSize'].sum()/
 ↪ziplevelRows['inSize'].sum())
        assert(ziplevelRows.size>1000)

        dictData = { 'dataset' : ziplevelRows['dataset'].iloc[0] ,
                   'zipcmd':  ziplevelRows['zipcmd'].iloc[0] ,
                   'ziplevel':  ziplevelRows['ziplevel'].iloc[0] ,
                   'jobs':  ziplevelRows['jobs'].iloc[0] ,
                     'totalInTime' : ziplevelRows['inTime'].sum() ,
                   'totalInSize':  ziplevelRows['inSize'].sum() ,
                   'totalOutTime':  ziplevelRows['outTime'].sum() ,
                   'totalOutSize':  ziplevelRows['outSize'].sum() ,

                   }

        dictData['Compression'] = dictData['totalOutSize']/␣
 ↪dictData['totalInSize']
        dictData['inRateGbps'] = dictData['totalInSize']/␣
 ↪dictData['totalInTime'] * dictData['jobs'] *8/1e9
        dictData['outRateGbps'] = dictData['totalOutSize']/␣
 ↪dictData['totalOutTime']* dictData['jobs'] *8/1e9

        dataframeSum = dataframeSum.append(dictData, ignore_index=True)
```

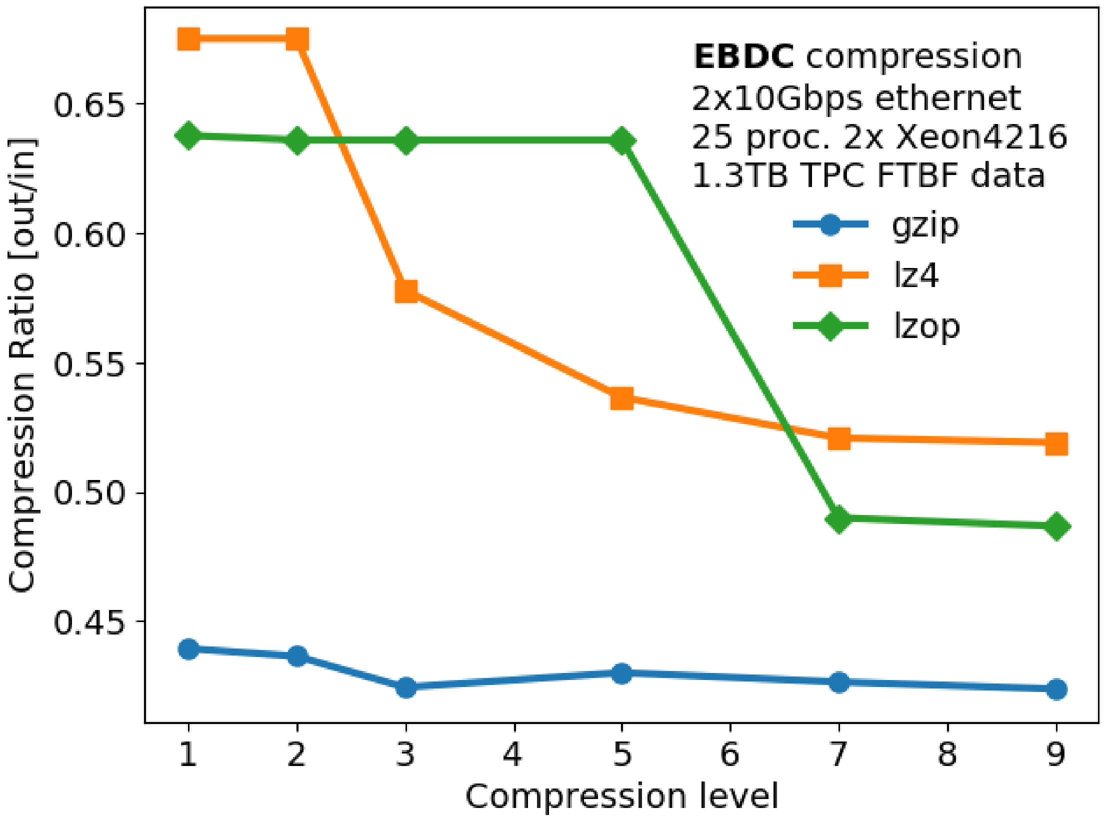
```
processing  gzip . 1  size=  2133 compression ratio =  0.43932139377897234
processing  gzip . 2  size=  2133 compression ratio =  0.43639353842830403
processing  gzip . 3  size=  2133 compression ratio =  0.424364599250152
processing  gzip . 5  size=  2133 compression ratio =  0.4300136029630121
processing  gzip . 7  size=  2133 compression ratio =  0.4264271430634125
processing  gzip . 9  size=  2133 compression ratio =  0.42370792929411943
processing  lz4 . 1  size=  2133 compression ratio =  0.6751259046982664
processing  lz4 . 2  size=  2133 compression ratio =  0.6751259046982664
processing  lz4 . 3  size=  2133 compression ratio =  0.5778558660340661
processing  lz4 . 5  size=  2133 compression ratio =  0.5365085644373812
processing  lz4 . 7  size=  2133 compression ratio =  0.5207546008082999
processing  lz4 . 9  size=  2133 compression ratio =  0.5189943751016245
processing  lzop . 1  size=  2133 compression ratio =  0.6377294848760965
processing  lzop . 2  size=  2133 compression ratio =  0.6359905949774498
processing  lzop . 3  size=  2133 compression ratio =  0.6359905949774498
processing  lzop . 5  size=  2133 compression ratio =  0.6359905949774498
processing  lzop . 7  size=  2133 compression ratio =  0.48989600087271923
processing  lzop . 9  size=  2133 compression ratio =  0.4866962781452209
```

[15]:
```python
import matplotlib.pyplot as plt
import numpy as np

Colors = ['#1f77b4',
          '#ff7f0e',
          '#2ca02c',
          '#d62728',
          '#9467bd',
          '#8c564b',
          '#e377c2',
          '#7f7f7f',
          '#bcbd22',
          '#17becf',
          '#1a55FF']
Markers = ['o' , 's',  'D','p' , 'P']

font = {'size'   : 14}
plt.rcdefaults()
plt.rc('font', **font)

studytitle_sup = studytitle + "\n{:d} proc. 2x Xeon4216\n{:.1f}TB TPC FTBF␣
 ↪data".format(
    dataframeSum['jobs'].iloc[0], dataframeSum['totalInSize'].iloc[0]/1e12)
```

### 4.1 Compression plot

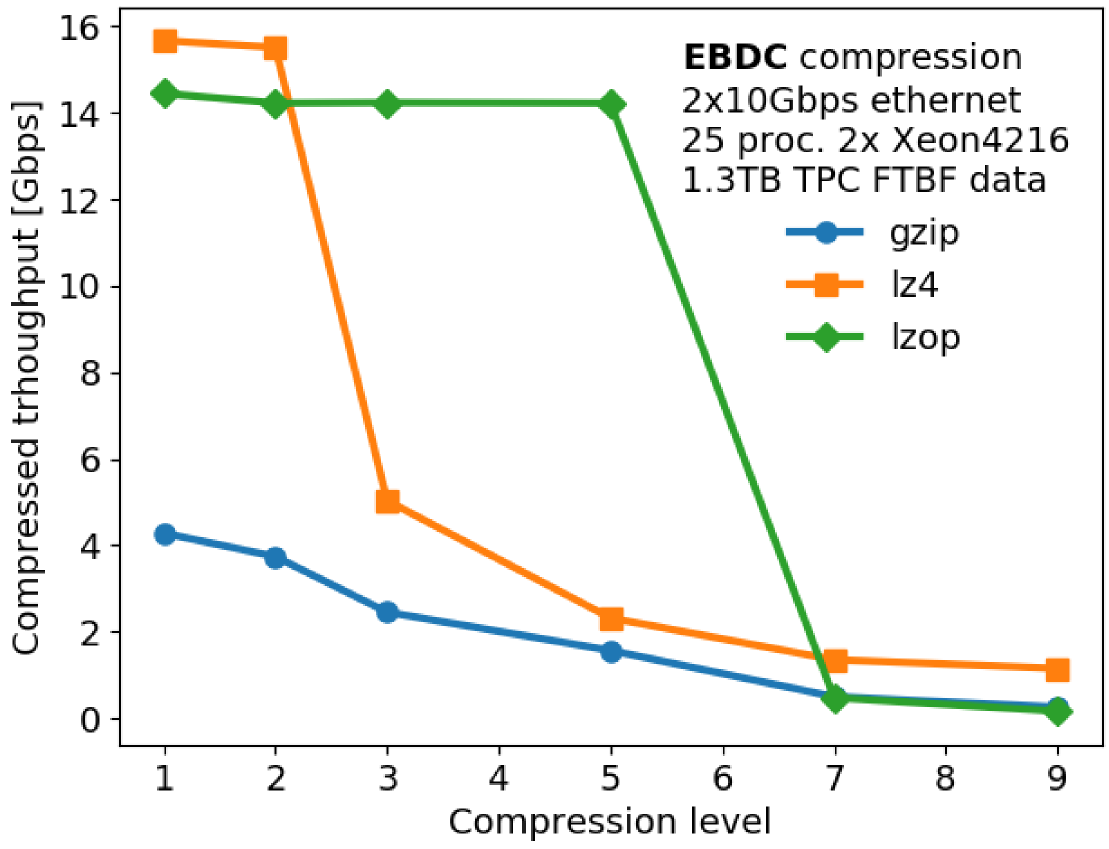
[16]:
```python
# dataframeSum.plot(x = 'ziplevel', y = "Compression")

fig = plt.figure()
ax = fig.add_axes([0.15, 0.15, 0.85, 0.85])
plt.xlabel('Compression level')
plt.ylabel('Compression Ratio [out/in]')

markiter = iter(Markers);
coleriter = iter(Colors);
```

```
for zipcmd in dataframeSum.zipcmd.unique():
    zipRows = dataframeSum.loc[dataframeSum['zipcmd'] == zipcmd]
    ax.plot(zipRows['ziplevel'].to_numpy(), zipRows['Compression'].to_numpy(),
            marker=next(markiter), color=next(coleriter), markersize = 8,␣
 ↪linewidth = 3,
            label=zipcmd)

plt.legend(loc='best',title = studytitle_sup, frameon=False)

plt.savefig(os.path.join(DataDir,"Compression.png"), dpi=150)
plt.savefig(os.path.join(DataDir,"Compression.pdf"), dpi=150)
```



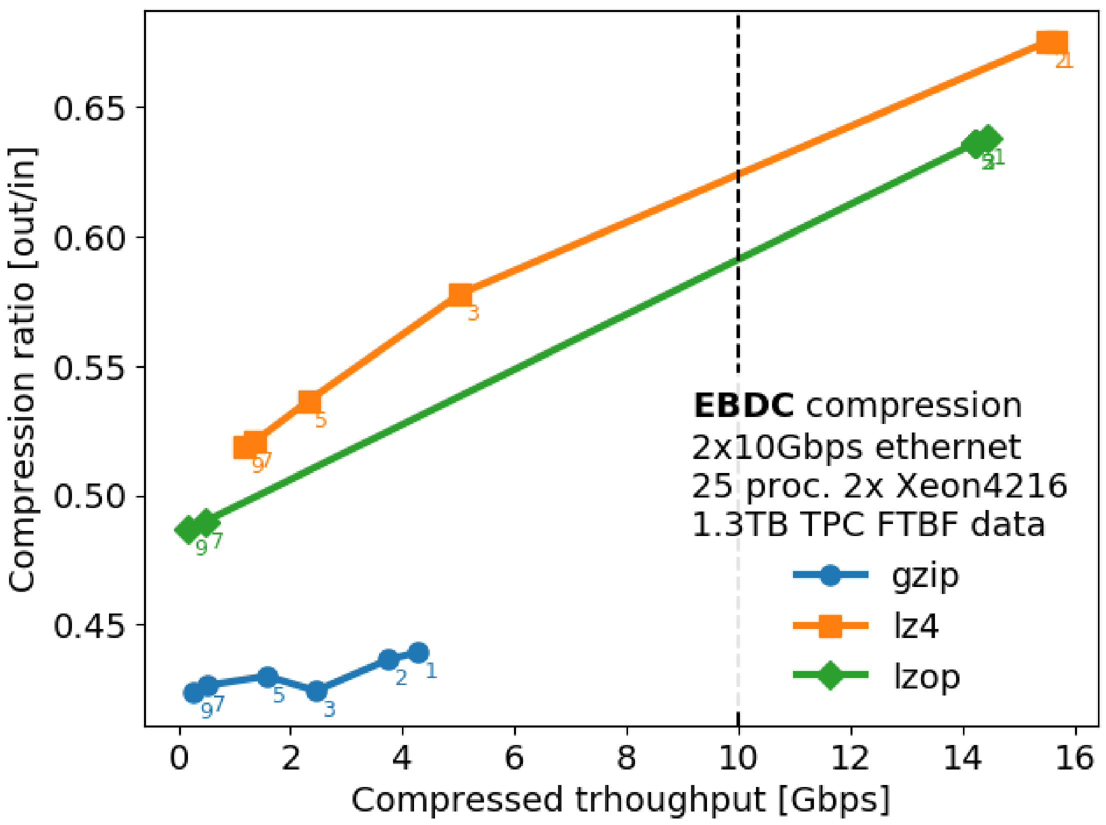
## 4.2 Compressed throughput

[17]:
```
# dataframeSum.plot(x = 'ziplevel', y = "Compression")

fig = plt.figure()
ax = fig.add_axes([0.15, 0.15, 0.85, 0.85])
plt.xlabel('Compression level')
plt.ylabel('Compressed trhoughput [Gbps]')

markiter = iter(Markers);
coleriter = iter(Colors);

for zipcmd in dataframeSum.zipcmd.unique():
    zipRows = dataframeSum.loc[dataframeSum['zipcmd'] == zipcmd]
    ax.plot(zipRows['ziplevel'].to_numpy(), zipRows['outRateGbps'].to_numpy(),
            marker=next(markiter), color=next(coleriter), markersize = 8,␣
 ↪linewidth = 3,
            label=zipcmd)

plt.legend(loc='best',title = studytitle_sup, frameon=False)

plt.savefig(os.path.join(DataDir,"Throughput.png"), dpi=150)
plt.savefig(os.path.join(DataDir,"Throughput.pdf"), dpi=150)
```



### 4.3 Work point curve

[18]:
```
# dataframeSum.plot(x = 'ziplevel', y = "Compression")

fig = plt.figure()
ax = fig.add_axes([0.15, 0.15, 0.85, 0.85])
plt.ylabel('Compression ratio [out/in]')
plt.xlabel('Compressed trhoughput [Gbps]')

markiter = iter(Markers);
coleriter = iter(Colors);

for zipcmd in dataframeSum.zipcmd.unique():
    zipRows = dataframeSum.loc[dataframeSum['zipcmd'] == zipcmd]
    outRateGbps = zipRows['outRateGbps'].to_numpy()
    Compression = zipRows['Compression'].to_numpy()
    ziplevel = zipRows['ziplevel'].to_numpy()
    c = next(coleriter)
    ax.plot(outRateGbps, Compression,
            marker=next(markiter), color=c, markersize = 8, linewidth = 3,
            label=zipcmd)

    for i in range(0, len(outRateGbps)):
        plt.text(outRateGbps[i]+.1, Compression[i]-.01,str(ziplevel[i]),␣
 ↪fontsize=9, color=c)

#              [str(i) for i in zipRows['ziplevel'].to_numpy()], fontsize=9)

ax.add_line(plt.Line2D([10, 10],ax.get_ylim(), color = 'black', linestyle =␣
 ↪'--'))

plt.legend(loc='best',title = studytitle_sup,
           edgecolor = 'white', frameon=True, facecolor='white', framealpha=0.9)

plt.savefig(os.path.join(DataDir,"FOM.png"), dpi=150)
plt.savefig(os.path.join(DataDir,"FOM.pdf"), dpi=150)
```



## 5 Scratch

[19]:
```
# %save_html os.path.join(DataDir,"analysis.html")
import sys
from subprocess import check_call

d, fname = os.path.split(sys.executable)
# print (d,fname)
check_call([os.path.join(d,'ipython'), 'nbconvert',
            '--to', 'html',
            'RateAnalysis.ipynb',
            '--output',
            os.path.join(DataDir,"analysis.html")])
check_call([os.path.join(d,'ipython'), 'nbconvert',
            '--to', 'pdf',
            'RateAnalysis.ipynb',
            '--output',
            os.path.join(DataDir,"analysis.pdf")])
```

[19]:
```
0
```

[20]:
```
dataframeSum
```

[20]:
```
      dataset zipcmd  ziplevel  jobs   totalInTime   totalInSize  totalOutTime  \
0   gzip-1-25   gzip         1    25    27604.6275  1.345599e+12    27605.2246
1   gzip-2-25   gzip         2    25    31389.4314  1.345599e+12    31390.0466
2   gzip-3-25   gzip         3    25    46470.5689  1.345599e+12    46471.4053
3   gzip-5-25   gzip         5    25    73536.7874  1.345599e+12    73537.9159
4   gzip-7-25   gzip         7    25   225710.8268  1.345599e+12   225713.7803
5   gzip-9-25   gzip         9    25   432297.9196  1.345599e+12   432303.2072
6    lz4-1-25    lz4         1    25    11604.5203  1.345599e+12    11607.9046
7    lz4-2-25    lz4         2    25    11723.1113  1.345599e+12    11726.4314
8    lz4-3-25    lz4         3    25    30892.6923  1.345599e+12    30902.6711
9    lz4-5-25    lz4         5    25    62124.8162  1.345599e+12    62144.7757
10   lz4-7-25    lz4         7    25   102965.9950  1.345599e+12   102999.0713
11   lz4-9-25    lz4         9    25   120272.7466  1.345599e+12   120311.7069
12  lzop-1-25   lzop         1    25    11890.7268  1.345599e+12    11891.3771
13  lzop-2-25   lzop         2    25    12042.1961  1.345599e+12    12042.8259
14  lzop-3-25   lzop         3    25    12034.8472  1.345599e+12    12035.4690
15  lzop-5-25   lzop         5    25    12042.5358  1.345599e+12    12043.1717
16  lzop-7-25   lzop         7    25   273029.2722  1.345599e+12   273037.1193
17  lzop-9-25   lzop         9    25   752188.2002  1.345599e+12   752208.8112

    totalOutSize  Compression  inRateGbps  outRateGbps
0   5.911506e+11     0.439321    9.749086     4.282889
1   5.872109e+11     0.436394    8.573582     3.741383
2   5.710247e+11     0.424365    5.791190     2.457532
3   5.786260e+11     0.430014    3.659663     1.573681
4   5.738001e+11     0.426427    1.192322     0.508432
5   5.701411e+11     0.423708    0.622533     0.263769
6   9.084490e+11     0.675126   23.190952    15.652248
7   9.084490e+11     0.675126   22.956353    15.494040
8   7.775625e+11     0.577856    8.711441     5.032332
9   7.219256e+11     0.536509    4.331922     2.323367
10  7.007271e+11     0.520755    2.613677     1.360647
11  6.983585e+11     0.518994    2.237580     1.160915
12  8.581284e+11     0.637729   22.632752    14.432784
13  8.557885e+11     0.635991   22.348073    14.212421
14  8.557885e+11     0.635991   22.361719    14.221108
15  8.557885e+11     0.635991   22.347442    14.212013
16  6.592038e+11     0.489896    0.985681     0.482867
17  6.548982e+11     0.486696    0.357783     0.174127
```

[ ]: