RateAnalysis

September 15, 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 TPC connection to multiple `ncat` processes either at `localhost` or remote which can be started with start-sink.sh

2 Inputs

[1]:
```
# DataDir = './data/'
DataDir = './data_25x_localhost/'
studytitle = r"$\bf{EBDC}$" + " compression\nlocalhost loopback"
```

[2]:
```
# %matplotlib widget
# %matplotlib ipympl
%matplotlib inline
# well the html export like dump formats
```

3 Processing

[3]:
```
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))
```

```
skip .ipynb_checkpoints
processing ./data_25x_localhost/gzip-1-25, gzip level1 x25
processing ./data_25x_localhost/gzip-2-25, gzip level2 x25
processing ./data_25x_localhost/gzip-3-25, gzip level3 x25
processing ./data_25x_localhost/gzip-5-25, gzip level5 x25
processing ./data_25x_localhost/gzip-7-25, gzip level7 x25
processing ./data_25x_localhost/gzip-9-25, gzip level9 x25
processing ./data_25x_localhost/lz4-1-25, lz4 level1 x25
processing ./data_25x_localhost/lz4-2-25, lz4 level2 x25
processing ./data_25x_localhost/lz4-3-25, lz4 level3 x25
processing ./data_25x_localhost/lz4-5-25, lz4 level5 x25
processing ./data_25x_localhost/lz4-7-25, lz4 level7 x25
processing ./data_25x_localhost/lz4-9-25, lz4 level9 x25
processing ./data_25x_localhost/lzop-1-25, lzop level1 x25
processing ./data_25x_localhost/lzop-2-25, lzop level2 x25
processing ./data_25x_localhost/lzop-3-25, lzop level3 x25
processing ./data_25x_localhost/lzop-5-25, lzop level5 x25
processing ./data_25x_localhost/lzop-7-25, lzop level7 x25
processing ./data_25x_localhost/lzop-9-25, lzop level9 x25
```

4 Plot

[4]:
```
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)
```

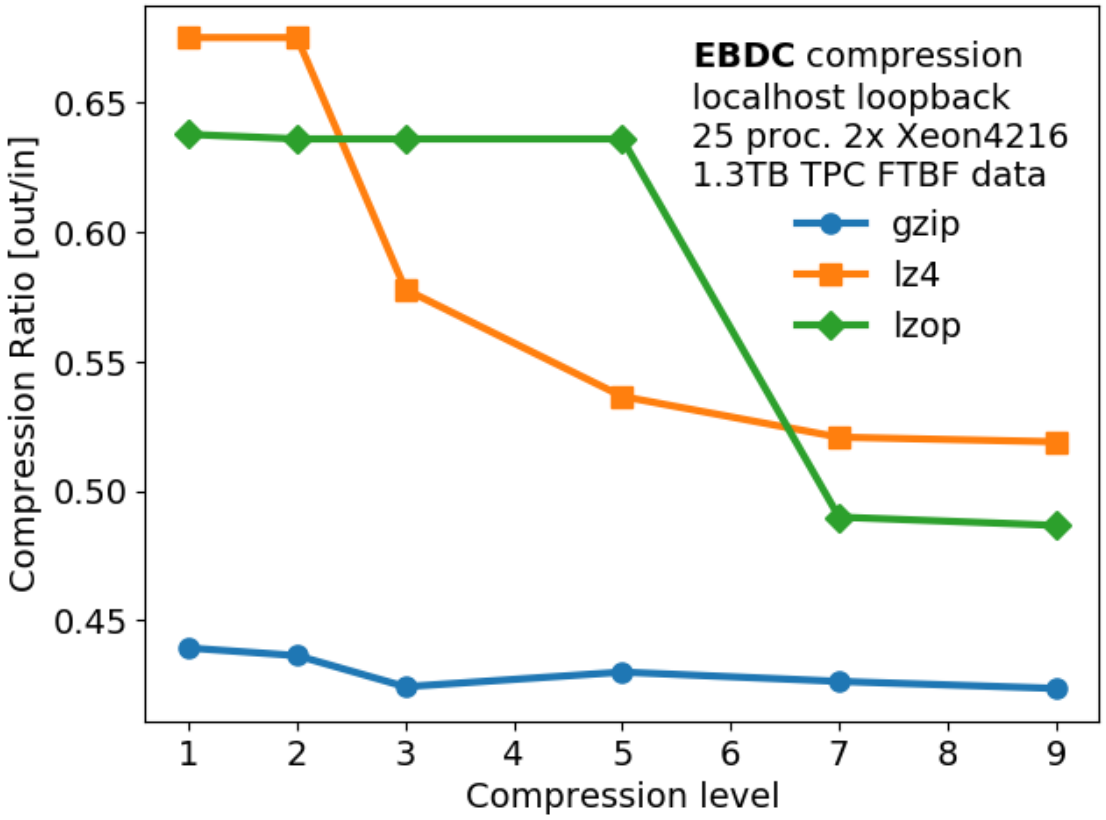
```
processing  gzip . 1  size=  2133 compression ratio =  0.43932139377897234
processing  gzip . 2  size=  2133 compression ratio =  0.43639353842830403
processing  gzip . 3  size=  2133 compression ratio =  0.424364599250152
processing  gzip . 5  size=  2133 compression ratio =  0.4300136029630121
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
```

[5]:
```
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

[6]:
```
# 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)
```

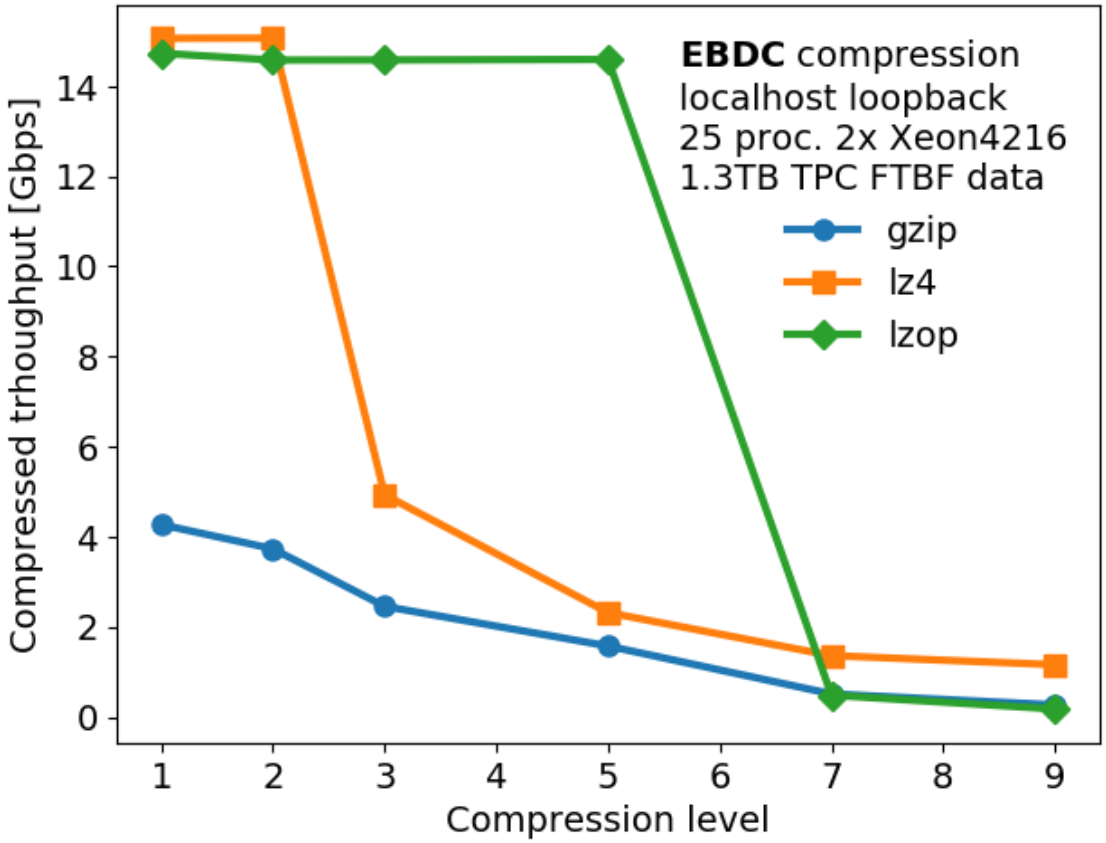
```
plt.savefig(os.path.join(DataDir,"Compression.pdf"), dpi=150)
```



4.2 Compressed throughput

[7]:
```
# 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)
```

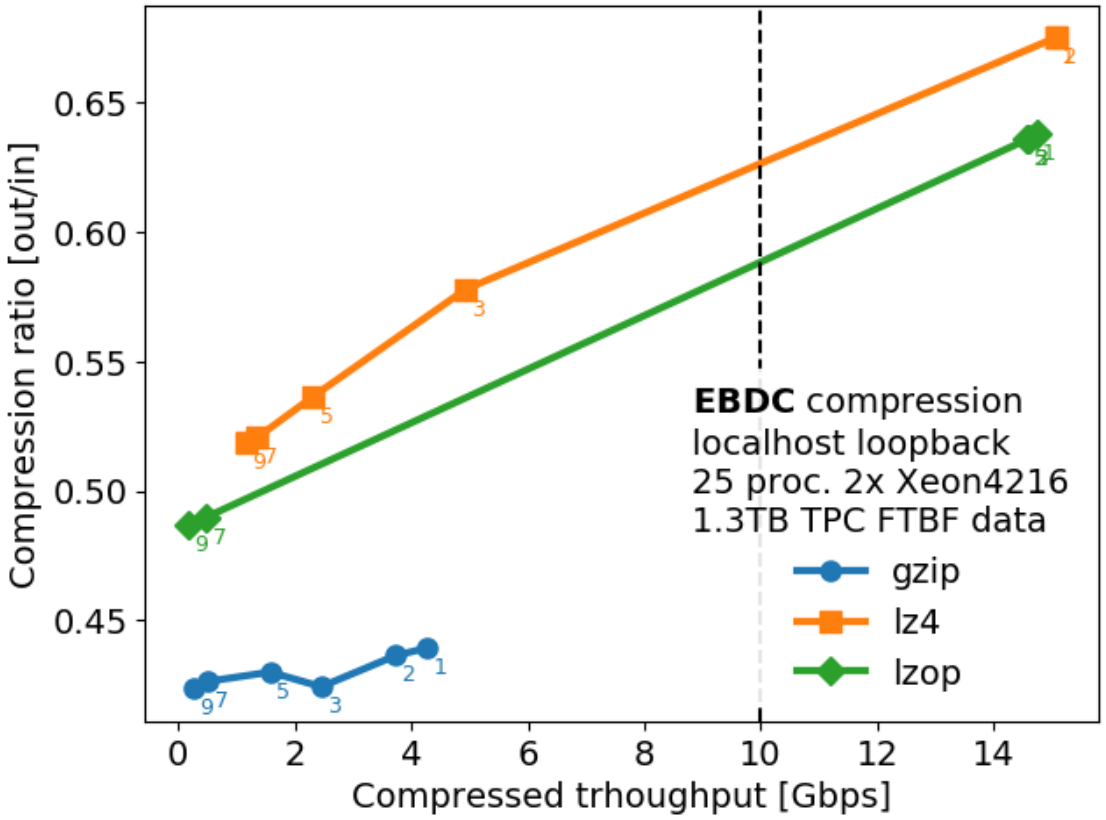
```
plt.legend(loc='best',title = studytitle_sup, frameon=False)

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

[8]:
```
# 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

[9]:
```
# %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")])
```

```
/home/jinhuang/anaconda3/bin python
```

[9]:
```
0
```

[10]:
```
dataframe.loc[dataframe['jobID'] == '108']
```

[10]:

	dataset	zipcmd	ziplevel	jobs	jobID	inTime	inSize	\
10	gzip-1-25	gzip	1	25	108	216.4719	1.281674e+10	
247	gzip-2-25	gzip	2	25	108	240.3604	1.281674e+10	
484	gzip-3-25	gzip	3	25	108	335.3750	1.281674e+10	
721	gzip-5-25	gzip	5	25	108	518.5007	1.281674e+10	
958	gzip-7-25	gzip	7	25	108	1447.6632	1.281674e+10	
1195	gzip-9-25	gzip	9	25	108	2831.0358	1.281674e+10	
1432	lz4-1-25	lz4	1	25	108	97.3737	1.281674e+10	
1669	lz4-2-25	lz4	2	25	108	101.2019	1.281674e+10	
1906	lz4-3-25	lz4	3	25	108	237.2366	1.281674e+10	
2143	lz4-5-25	lz4	5	25	108	429.3463	1.281674e+10	
2380	lz4-7-25	lz4	7	25	108	770.7569	1.281674e+10	
2617	lz4-9-25	lz4	9	25	108	1038.7559	1.281674e+10	
2854	lzop-1-25	lzop	1	25	108	96.6475	1.281674e+10	
3091	lzop-2-25	lzop	2	25	108	88.0183	1.281674e+10	
3328	lzop-3-25	lzop	3	25	108	96.5301	1.281674e+10	
3565	lzop-5-25	lzop	5	25	108	88.5881	1.281674e+10	
3802	lzop-7-25	lzop	7	25	108	1830.3287	1.281674e+10	
4039	lzop-9-25	lzop	9	25	108	5201.6103	1.281674e+10	

```
        outTime       outSize
10     216.4731  4.159831e+09
247    240.3624  4.107418e+09
484    335.3776  3.916616e+09
721    518.5056  3.884543e+09
958   1447.6791  3.846747e+09
1195  2831.0647  3.837098e+09
1432    97.3937  6.133746e+09
1669   101.2205  6.133746e+09
1906   237.3043  5.216794e+09
2143   429.4768  4.760619e+09
2380   770.9690  4.558153e+09
2617  1039.0030  4.515718e+09
2854    96.6492  6.048131e+09
3091    88.0204  6.033408e+09
3328    96.5319  6.033408e+09
3565    88.5901  6.033408e+09
3802  1830.3778  4.465579e+09
4039  5201.7425  4.438163e+09
```

[ ]: