Identifying Anomalous File Transfer Events in LCLS Workflow

Mengying Yang, Xinyu Liu, Wilko Kroeger, John Wu, Alex Sim

Outline

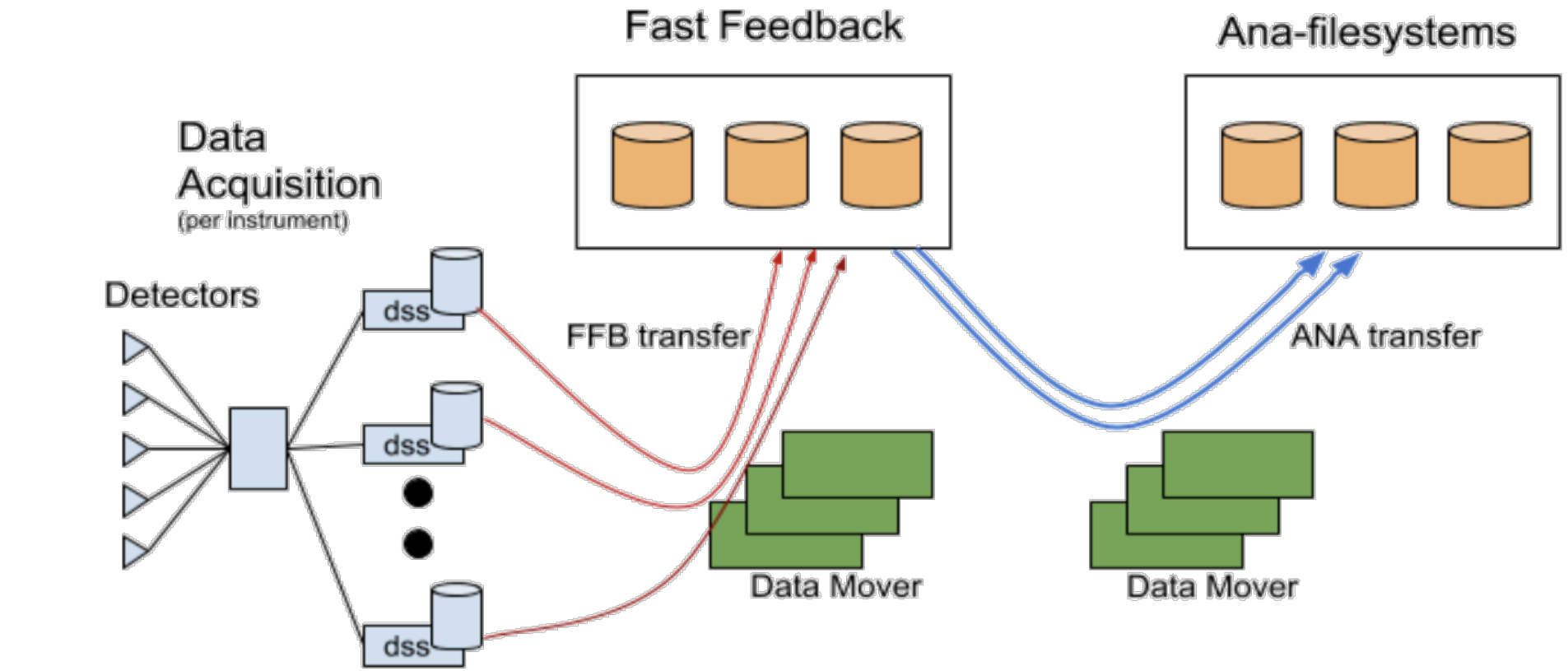
- Goal
- Background
- Data
- Exploratory Data Analysis
- Anomaly Detection Algorithms
- Result
- Summary and Challenges

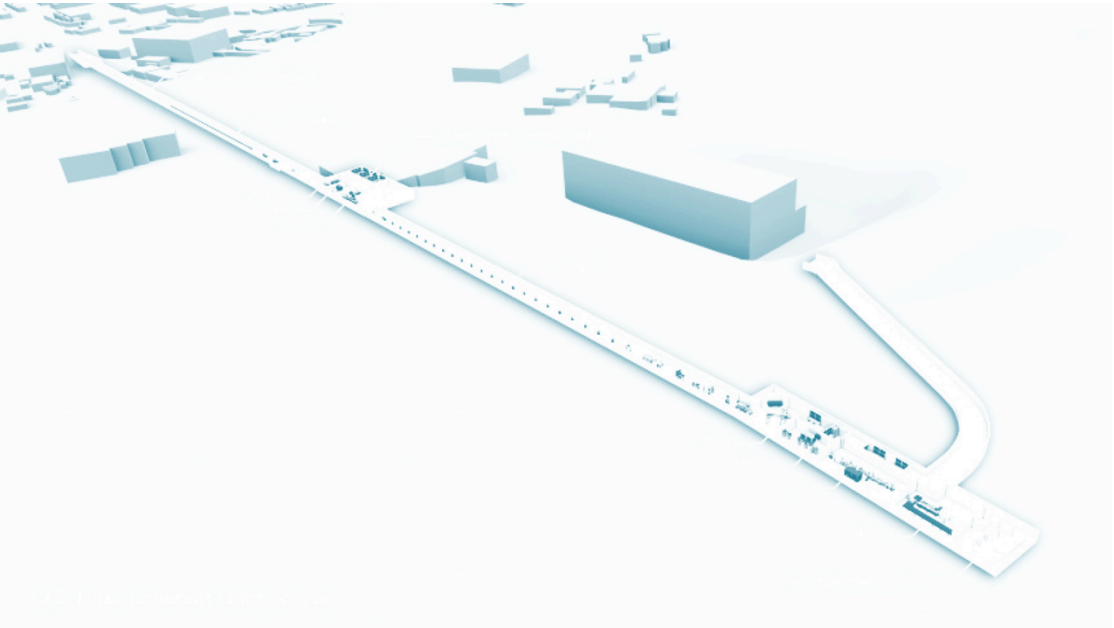
LCLS: Linac Coherent Light Source (at SLAC)

X-Swap Extreme-Scale Scientific Workflow Analysis and Prediction

Goal

- **Context**: Improve the "smartness" of large experimental facilities
  - Monitoring components
  - React to unusual events
- **Specific objective**: Help infrastructure operators to monitor the health of the system and anticipate potential failures for a large scientific facility known as Linac Coherent Light Source (LCLS)
  - Identify anomalous file transfers
  - Explore the significant variables will influence or correlate with anomalous file transfers





Background: LCLS

Data

- The LCLS file transfer dataset collected by SLAC National Accelerator Laboratory, and includes file transferred information through the Fast-Feedback system(FFB) and analysis storage system(ANA) from May 2017 to Jan 2018
- The dataset contains 258,765 observations with 10 variables.
    - The FFB transfers accounts for 131,274 observations and the ANA transfer for 127,491 observations.
    - The variables used in our study are: the start and stop time of a file transfer (epoch time in seconds); file transfer rate (MiB/sec), file size (gigabytes); a boolean variable of whether it is an FFB or ANA transfer.
    - Other variables: the name of a file, LCLS instrument the data was collected with, file system the data were written to, and more

Data Distribution Exploration

STATS	size (gigabyte)	transfer rate(MB/s)
median	4.0	47.9
mean	13.8	81.2
std	22.1	91.0
min	0	0
max	1304.14	498.2

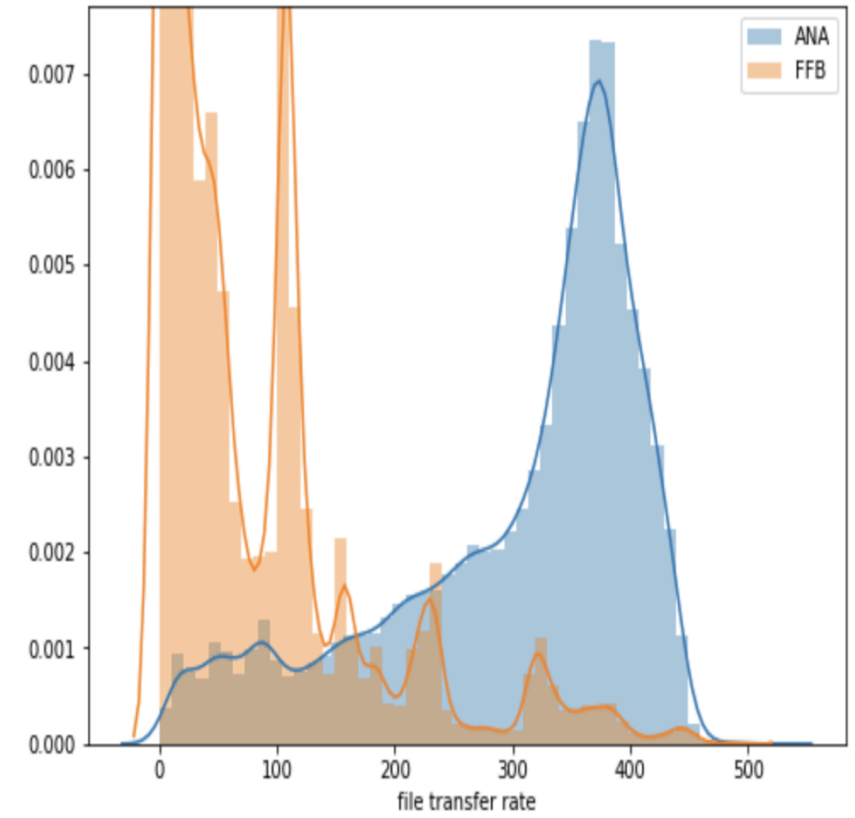
(a)

STATS	size (gigabyte)	transfer time (MB/s)
median	4.0	342.7
mean	13.29	301.8
std	21.49	109.7
min	0	0
max	1304.14	522.2

(b)

Table 1: (a)DSS to FFB (b)FFB to ANA

File transfer rate distribution for ANA and FFB



Unusual Data Behavior Exploration

- File Size
  - Unexpected large files
  - Zero file size

- File Transfer Rate
  - Slow transfer rate.

Unexpected Large Files

12 files (6 files in each dataset) whose file size is much larger than the 100 GiB that is enforced by the data acquisition control.

Happened within the same day -- likely due to a configuration error.

Removed these files in our analysis.

STATS	size (gigabyte)	transfer rate(MB/s)
median	4.0	47.9
mean	13.8	81.2
std	22.1	91.0
min	0	0
max	1304.14	498.2

(a)

STATS	size (gigabyte)	transfer time (MB/s)
median	4.0	342.7
mean	13.29	301.8
std	21.49	109.7
min	0	0
max	1304.14	522.2

(b)

Table 1: (a)DSS to FFB (b)FFB to ANA

Zero-Size Files

- 76 files with zero file size and zero transfer rate
- These files were likely produced by some failures in the data acquisition system

Slow Transfer Rate

- Definition:
  - File size larger than 1GB
  - The transfer rate that is lower than 1 percentile of all the transfer rates for files being transferred to FFB or ANA respectively
- Description:
  - Threshold for FFB and ANA are 4.3MB/s and 45.7MB/s respectively.
  - There are 912 and 872 observations of the very slow transfer rate data from FFB and ANA respectively.

Anomaly Detection Methods

- Model-Based Detection Method
- Distribution-Based Detection Method

Model-Based Detection Method

- Predict Base line
  - Applied $\log_2$ transformation of both size and transfer time
  - B-spline minimum Quantile Regression
- Check whether the error percentage between predicted time and actual transfer time is above certain threshold
  - Error percentage is simply the error divided by the predicted base time
  - Threshold is q-percentile of error percentage
- Refitted the B-spline Quantile regression at every segment with new data appended to the old

Model-Based Detection Method

Algorithm 1 Model-Based Detection

$y\{0j\} \leftarrow$ log transfer time for file j *for first a files*
$s\{0j\} \leftarrow$ log transfer size for file j for first a files
Apply BSpline quantile regression with $s\{0\}$ as covariate and $y\{0\}$ as response to get $\hat{y}_0$
$e\{0\} \leftarrow (2^{y\{0\}} - 2^{\hat{y}_0})/2^{\hat{y}_0}$ j *for first a files*
for *each segment i of length h* **do**
  $y\{ij\} \leftarrow$ log transfer time for file j in Segment i
  $s\{ij\} \leftarrow$ log transfer size for files j in Segment i
  Apply BSpline quantile regression to estimate to get $\hat{y}\{ij\}$in Segment i
  $e\{ij\} \leftarrow 2^{y\{ij\}} - 2^{\hat{y}\{ij\}}/2^{\hat{y}\{ij\}}$ j in Segment i
  Append $e\{ij\}toe\{0\}$
  $threshold \leftarrow q$ *percent quantile of e{0}*
  if $e\{ij\} \geq threshold$ **then**
    State file i as an anomaly
  end if
end for

Predict time from file sizes

Distribution-Based Detection Method

- Motivation
  - The slope of the baseline can be interpreted as the inverse of the maximum file transfer rate
  - The distribution of the error percentage is highly correlated with the distribution of the file transfer rate.

Distribution-Based Detection Method

- Uses the file transfer rate distribution of the past 5000 file transfers and sets the 0.2 percentile as the threshold

- Keep tracking the mean of the 5000 transfer rates. When adding a new observation increases the new mean by more than the 75 percentile of the mean difference distribution, we set this observation rate as the mean of the all threshold

Algorithm 2 Distribution-Based Detection

data ← first m file transfer data
threshold{0} ← 0.1 percentile of first m points collected
mean{0} ← mean of the first m file transfer rate
data_copy ← first m file transfer rate data
for *each new file record i* **do**
  mean{*i*} ← mean of file transfer rate from record i-m to i
  dis{*i*} ← mean{i} - mean{i-1}
  if *dis*{*i*} ≥ (75*percentileofalldis*) **then**
    threshold{*i*} ← mean of all previous threshold
    data_copy{*i*} ← mean of all previous threshold
  else
    threshold{*i*} ← q percentile of data_copy from i-m to i
    data_copy{*i*} ← new file transfer rate i
  end if
  if *threshold*{*i*} ≥ *newfiletransferrate*{*i*} **then**
    State file *i* as an anomaly
  end if
end for

Detection based on file-transfer rates

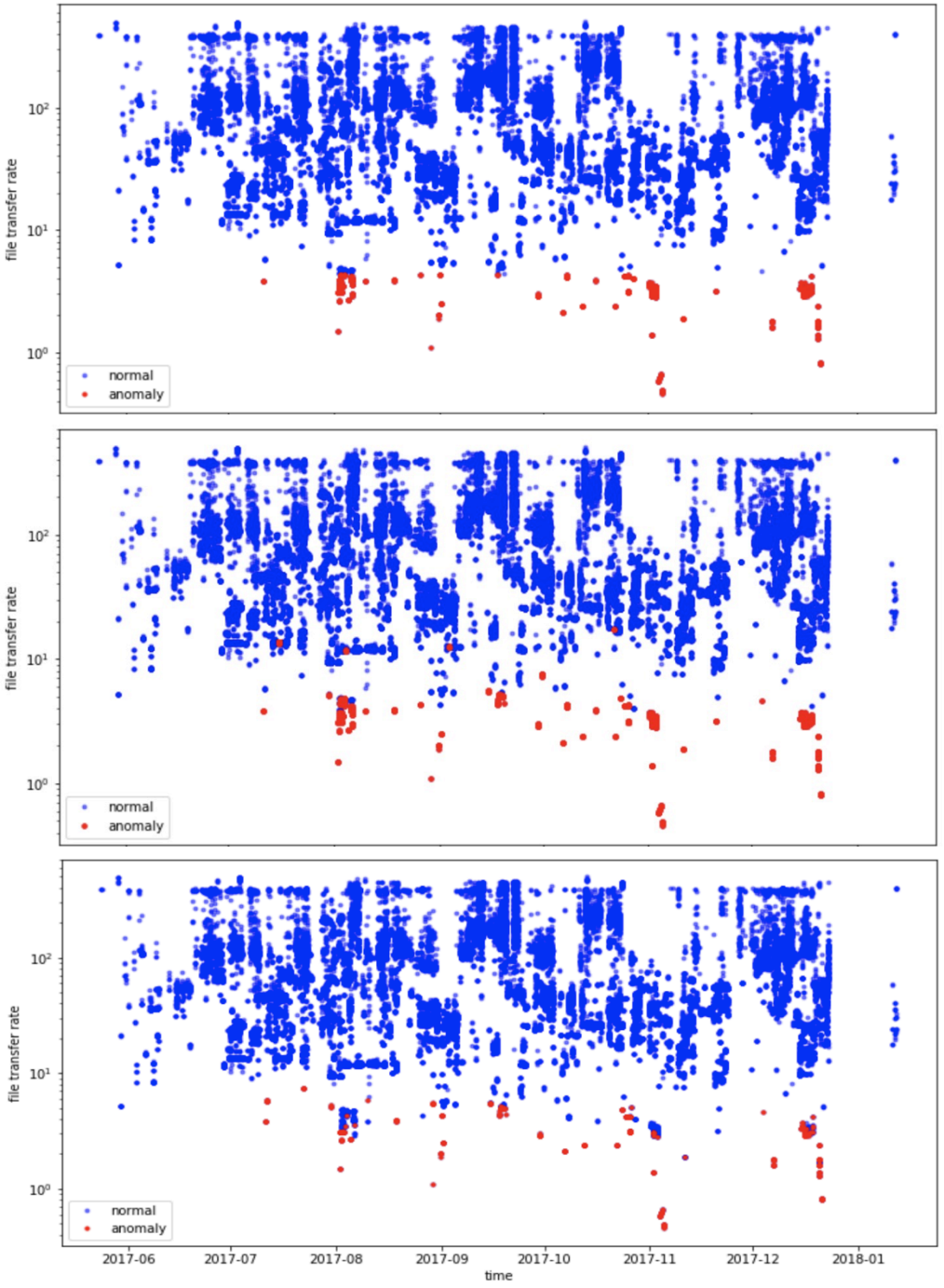
Results for FFB

-The y-axis represent file transfer rate and in the log scale

-The first plot shows the position of the actual slowest one percent rate.

-The second and third are the anomalies detected from the model-based method and distribution-based method respectively.

Visually, the plots are similar.



Results for FFB

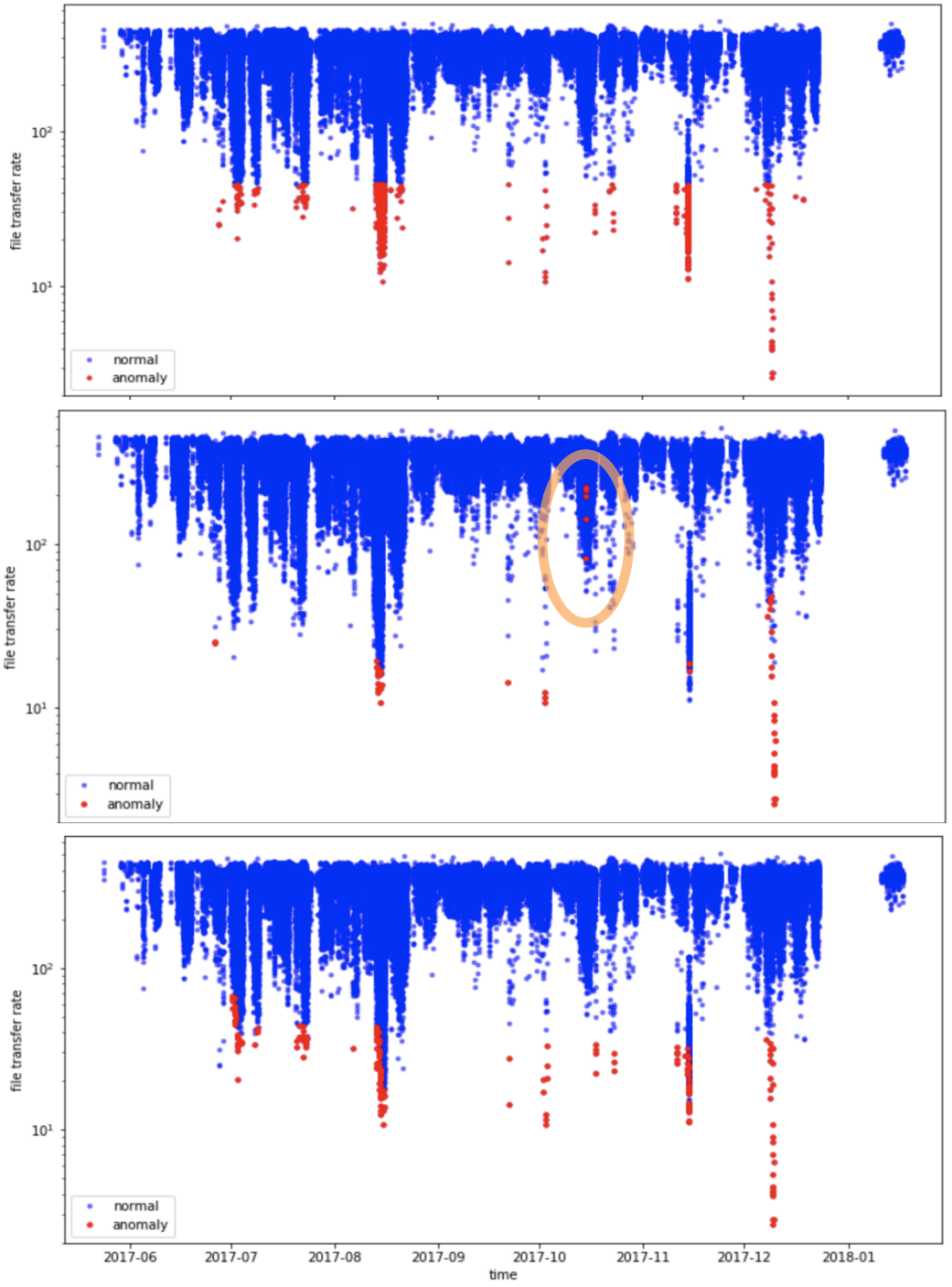
There are 2526 hours in total and 120 hours contain the slow transfer rates.

	Number of hour detected	Number of matched hours	Precision	Recall
Model-based detection	129	109	84.5% (=109/129)	90% (=109/120)
Distribution-based detection	71	64	90.14% (=64/71)	53.3% (=64/120)
True value	120	120		

Results for ANA

-The y-axis represent file transfer rate and in log scale
-The first plot shows the position of the actual slowest one percent rate.
-The second and third are the anomalies detected from the model-based method and distribution-based method respectively.

The plots are not identical to the result of FFB, especially for the anomalies detected by the model-based method.



Results for ANA

There are 2598 hours in total and 97 hours contain the slow transfer rates.

	Number of hour detected	Number of matched hours	Precision	Recall
Model-based detection	20	18	90% (=18/20)	18.6% (=18/97)
Distribution-based detection	50	47	94% (=47/50)	48.45% (=47/97)
True value	97	97		

Comparison

- Model-based method performs better for the FFB transfers and worse for the ANA transfers since this method works better when the slowest transfer rate is stable and the transfer behavior is unstable for consecutive transfers
- Distribution-based method performs better for files transfer to ANA since this dataset has a stable consecutive transfer behavior but unstable lowest rate
- Distribution-based method is computationally less expensive

Summary

- Our key objective is to identify unusual file transfers in the LCLS data system.
- The initial data exploration helped us identify files with zero size and unexpectedly large sizes
- We proposed two methods to detect slow transfers, one based on a performance model and another based on the observed distribution of file transfer rates
- From the tests, we observed that model-based method works better for transfers to FFB, while the distribution-based method works better for transfers to ANA

Challenges and Future Work

- Combine the two anomaly detection methods
- Incorporate with other variables
- Combine with other datasets, such as file system information over time, to investigate the reason that the anomaly transfer happened
- Require further statistical testing to determine that usefulness in detecting anomalous events.

Thank you!

JOHN'S EMAIL
JOHN.WU@NERSC.GOV