Network Performance Behavior Detection with Multivariate Clusters in Time Windows

Tyler Leibengood[1], Alina Lazar (advisor)[1], Alex Sim (advisor)[2], Kesheng Wu (advisor)[2]
[1]Youngstown State University, [2]Lawrence Berkeley National Laboratory

ABSTRACT

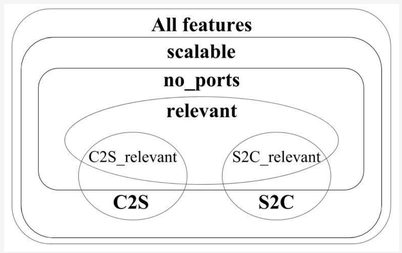
This project analyzes Tstat logs collected from Lawrence Berkeley National Laboratory's ESNet data transfer nodes to obtain insights on data transfer characteristics and behavior. Detecting anomalies in network transfers at the package level will provide solutions for improving network transfer rate. Several feature subsets from the Tstat logs were identify as good predictors for low network throughput. Dimensionality reduction was used to reduce the number of features and to select several sets of prominent features. K-means clustering provided a way to group data transfers by transfer quality. T-SNE was used to visualize and verify multi-variate clustering results in two-dimensional scatter plots. The results indicate that there is high correlation between the percentage of the smallest cluster of transfers and average throughput per time window for low throughput.

DATA

- Tstat data measures TCP flows
- Contains 104 features
- Data used from 5 out of 8 DTNs at NERSC
- Time collected: 5/1/2017 - 5/30/2017

METHODS

- **PCA** is used to select **5 prominent features** from 10 feature subsets



- **K-Means** is used on **prominent features** in 30 minute windows with 2 clusters:
  - Normal and Anomalous
- The percentage of flows in the smaller cluster is compared to average throughput for every window.
- When the Normal cluster is exceptionally small, throughput should be observed as unusually low.

RESEARCH QUESTION

Can clustering multivariate data transfer measurements detect low network performance?

RESULTS

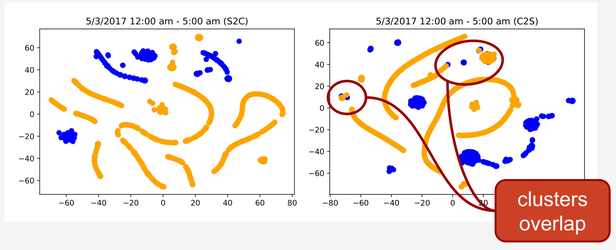
Feature Selection

Feat.	*S2C*	*S2C Recommended*	*C2S relevant*	*no_ports*
1	s_port:16	s_mss_min:89	c_cwin_max:70	c_mss:64
2	s_mss:87	s_pkts_dup:99	c_mss_max:65	c_mss_max:65
3	s_rst_cnt:18	s_cwin_min:94	con_t:42	s_mss:87
4	s_win_scl:84	s_mss_max:88	c_win_min:68	c_win_scl:61
5	s_mss_min:89	s_win_min:91	c_rst_cnt:4	s_rst_cnt:18

This table lists the set names and the PCA selected features belonging to each set. Repeated features are outlined in matching colored boxes.

- 4/10 of the PCA selected feature sets consistently detect abnormally low throughput.

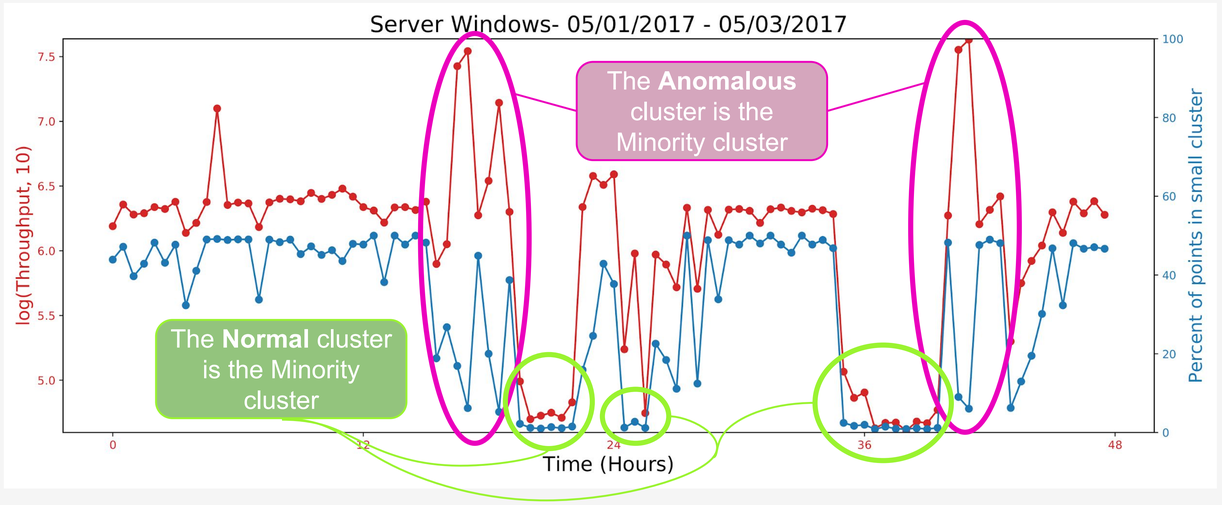
Clustering Verification



The clustering of 10 adjacent windows is visualized to determine if clustering is consistent between windows. For successful clustering, the labeled clusters should be separated.

- The *S2C* graph shows accurate classification.
- The *C2S* graph shows inaccurate classification.

Network Behavior Detection



This figure shows the comparison of percent data transfers in the Minority cluster with the average throughput for 30 minute windows in the first two days of data. When there are more Anomalous transfers than Normal transfers, low throughput is detected accurately. Because the detection system doesn't identify the Normal cluster, cases when there are more Normal transfers than Anomalous transfers are represented poorly.

RESULTS

Sequence Error

- Log(Throughput) is normalized
- **Root Mean Squared Error** compares the model to log(Throughput) for Throughput less than 1 Mb/sec.

	S2C	*S2C Recommended*	*C2S relevant*	*no_ports*
RMS Error	0.2057	0.1962	0.2661	0.2035

C2S relevant has the worst RMS Error and *S2C* has the best. The range in the RMS Error is 0.0699 and the average is 0.2179. The overall range is only 32% of the average RMS Error. Each feature subset has a relatively similar RMS Error.

- Root Mean Squared Error could be minimized by finding a more direct relationship with throughput and aligning data

CONCLUSIONS

- Multivariate clustering methods can be used to detect low performance in a data transfer network.
- We identified feature sets in Tstat data that accurately cluster Normal and Anomalous data transfers.
- There is high correlation between the percentage of normal transfers and throughput.
- In the future, we plan to modify our method to identify normal data transfers more accurately.

ACKNOWLEDGMENTS

The work on this project was completed by participation of the Berkeley Lab Undergraduate Faculty Fellowship (BLUFF) Program, managed by Workforce Development & Education at Berkeley Lab. This work was supported by the U.S. Department of Energy, under Contract No. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center. A special thanks goes to Mariam Kiran for advice on methods.