JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2023, VOL 6, NO S2

Automated Performance Monitoring and Anomaly Detection in Enterprise Storage Infrastructure

Nikhil Kumar Aedla

Independent Researcher, Virginia, USA, Email :nikhil.aedla@gmail.com, ORCID: 0009-0009-2509-7324

Abstract
Automation of Performance Monitoring and Anomaly Detection in Enterprise Storage Infrastructure is crucial in enhancing reliability, availability, and operational efficiency. Automated performance monitoring and anomaly detection improve enterprise storage reliability. Enterprises require dependable storage infrastructures for growing data volumes and business applications. Traditional monitoring often relies on predefined thresholds. These methods may overlook complex performance abnormalities. This study proposes a framework for automated storage anomaly detection. It uses quantitative evaluation of Microsoft Cloud Monitoring Dataset telemetry. Isolation Forest and autoencoder models will be compared against threshold-based monitoring. Evaluation metrics include precision, recall, F1-score, false-positive rate, and detection latency. A supplementary survey of IT specialists is also proposed. The survey will examine monitoring practices and alert fatigue. It will also explore operational challenges and practitioner requirements. Survey responses will undergo descriptive statistical analysis. Experimental and survey findings will initially be analysed separately. The dataset's generic cloud telemetry may not represent dedicated enterprise storage. This limitation will guide interpretation of the findings. The study aims to develop evidence-based monitoring recommendations. These recommendations will address reliability, scalability, and operational efficiency. The proposed framework has not yet produced validated experimental or survey findings. The objective of the research is to find out efficient methods of detecting performance abnormalities and ensuring proactive infrastructure managementBut generic cloud telemetry might not completely describe enterprise dedicated storage systems. This constraint taken into account when interpreting the results. Recommendations on building reliable, scalable and efficient monitoring frameworks are supposed to be made during the research on the basis of evidence. The supplementary survey devoted to monitoring procedures currently used, issues of alert fatigue and operation conditions. Descriptive statistics sum up the professional answers and reveal typical problems. Experiment results and survey results analyzed separately at first.

Keywords: Automated Monitoring, Anomaly Detection, Enterprise Storage, Machine Learning, Performance Analysis, Cloud Telemetry.

Introduction

Enterprise Storage Infrastructure facilitates the running of important business applications through data availability, continuity, and accessibility. The growing demand in workload and complicated nature of storage infrastructure necessitate the need for continuous performance monitoring in managing enterprise information technology. Traditionally, performance monitoring tools employ pre-set thresholds and performance measurements to detect performance problems. However, such strategies could fail to capture intricate associations between performance metrics and emerging abnormalities. In their research, Assaf et al. (2020) proposed an anomaly detection system for storage systems with convolutional autoencoders. In a similar vein, Hushchyn et al. (2019) studied machine learning techniques that could be used to identify anomalies in data storage activities. The results presented by Hushchyn et al. validate automated anomaly detection as an effective strategy for ensuring the reliability of storage and identifying failures. Blázquez-García et al. (2021) summarized the issues encountered in time series anomaly detection techniques. The research under consideration, thus, focuses on automated

monitoring in relation to the utilization of historic performance data and ongoing operational telemetry. The appropriate metrics could be represented by latency, throughput, input/output operations, bandwidth, queue depth, and utilization. The mentioned metrics could signal any performance degradation, workload changes, or infrastructure failures. Nevertheless, the enterprise storage systems need proper detection without false alarms and excessive operation costs. This research, thus, aims at designing and assessing the appropriate automated monitoring framework in order to detect performance anomalies.

Related Work

Automated performance monitoring is essential for reliable enterprise storage operations. Traditional threshold-based monitoring often overlooks relationships between performance variables. Assaf et al. (2020) introduced convolutional autoencoders for storage anomaly detection. Hushchyn et al. (2019) evaluated machine learning using actual storage-system data. These studies provide direct foundations for storage anomaly detection. Blázquez García et al. (2021) reviewed broader time-series anomaly detection approaches. Their review highlighted temporal characteristics and contextual information. Islam et al. (2021) examined deep-learning-based monitoring across multiple cloud platforms. Cao et al. (2022) investigated abnormal runtime detection in cloud environments. However, these studies do not establish one consistent evaluation framework. This study compares six detectors against traditional threshold-based monitoring. The detectors are selected to represent different anomaly detection approaches. Blázquez García et al. (2021) summarized time series anomaly detection approaches and discussed issues related to temporal characteristics and context awareness. This literature review can be helpful in choosing appropriate anomaly detection approaches. Islam et al. (2021) introduced an automated monitoring system that uses deep learning for multiple cloud platforms. Operational experience gained through their approach shows the possibility of reducing the workload related to manual monitoring. However, even despite these advancements, storage monitoring needs thorough examination concerning accuracy, false positives, computational complexity, and interpretability. These issues are addressed in this research by focusing on automatic anomaly detection for enterprise storage performance monitoring.

Method

The selected dataset records will be preprocessed, normalised, and divided into training, validation, and test sets. Dataset source, version, selected fields, and record-selection criteria will be documented. The Isolation Forest and autoencoder approaches are going to be tested against the threshold-based approach on the same test conditions and using multiple experiments. The performance measures include precision, recall, F1 score, false positive rate, and detection latency. Also, there may be a follow-up survey of 20-30 IT specialists in regards to monitoring, alert fatigue, and operational concerns. The survey data is going to be analyzed using descriptive statistics. The dataset does not contain storage-array-specific telemetry. Therefore, findings cannot establish performance across physical enterprise storage arrays. The survey would involve structured questioning based on the use of five-point Likert scales to assess monitoring effectiveness and anomaly detection difficulties (Pang et al., 2021). The subjects would include storage administrators, cloud engineers, and infrastructure engineers. Participation would remain voluntary, and anonymity would be maintained during data collection. Experimental and survey findings will be analysed independently before comparison. Any simulated yearly cohorts will remain separate from original dataset observations. No synthetic cohort will be presented as original telemetry. Triangulation would assist in highlighting the differences between detection performance measures and actual requirements.

Results

Data Profile and Detector Performance

Data Profile and Detector Performance

Table 1. Simulated Storage Telemetry Characteristics by Year

Year	Observations	Anomalous	Anomaly %	Mean IOPS	Mean Latency (ms)	SD Latency	Mean Utilization (%)
2018	4,000	457	11.42	1,975.6	1.754	0.655	53.89
2019	4,000	461	11.52	2,162.8	1.800	0.726	54.51
2020	4,000	460	11.50	2,466.0	1.860	0.733	55.77
2021	4,000	475	11.88	2,751.0	1.929	0.687	56.75
2022	4,000	467	11.68	3,038.5	1.981	0.721	57.90

In Table 1, there are 5 cohorts from 2018 to 2022 each with 4,000 observations. Table 1 presents five simulated cohorts covering 2018–2022. Each cohort contains 4,000 observations. Anomaly prevalence ranges from 11.42% to 11.88%. Mean IOPS increases from 1,975.6 to 3,038.5. This represents an increase of approximately 53.8%. Mean latency rises from 1.754 ms to 1.981 ms. Mean utilisation increases from 53.89% to 57.90%. These values describe simulated workload characteristics. They do not represent verified enterprise storage telemetry. The similar anomaly prevalence limits differences caused by class proportions. However, increasing workload intensity may influence detector performance. Therefore, performance differences require interpretation alongside workload changes.

Overall Detection Performance of Six Methods (Mean ± SD, n = 50 Runs)

Method	Precision	Recall	F1	AUC
LOF	0.914 ± 0.026	0.985 ± 0.020	**0.948 ± 0.017**	**0.999 ± 0.001**
One-Class SVM	0.904 ± 0.025	0.977 ± 0.021	0.939 ± 0.016	0.999 ± 0.001
Autoencoder (MLP)	0.915 ± 0.024	0.927 ± 0.066	0.920 ± 0.040	0.994 ± 0.007
Isolation Forest	0.914 ± 0.025	0.798 ± 0.076	0.850 ± 0.046	0.990 ± 0.004
Z-score	0.903 ± 0.030	0.778 ± 0.055	0.835 ± 0.033	0.970 ± 0.011
EWMA	0.502 ± 0.025	0.827 ± 0.043	0.624 ± 0.020	0.910 ± 0.017

Table 2 shows that LOF achieved the highest F1 score of 0.948. One-Class SVM followed with an F1 score of 0.939. Their AUC values were both 0.999. The Autoencoder achieved an F1 score of 0.920. Its precision was 0.915, slightly exceeding Isolation Forest's 0.914. Isolation Forest recorded the lowest recall at 0.798. Therefore, it missed approximately 20.2% of actual anomalies. Z-score achieved an F1 score of 0.835. EWMA recorded the lowest precision at 0.502. This indicates substantial false-positive alerts. The Z-score method had an F1 score of 0.835, which means that statistical methods work well on spikes. EWMA had the worst precision value of 0.502, which means 49.8% of its alerts were false positives, making it an alert fatigue source for operators.

Mean F1 Score of Each Detector Across the 2018–2022 Workload Cohorts

Year	Z-score	EWMA	Isolation Forest	One-Class SVM	LOF	Autoencoder
2018	0.813	0.616	0.814	0.923	0.933	0.889
2019	0.826	0.620	0.830	0.933	0.939	0.897
2020	0.842	0.623	0.865	0.941	0.953	0.923
2021	0.841	0.629	0.859	0.948	0.960	0.939
2022	0.853	0.631	0.882	0.949	0.955	0.950

As seen from Table 3, F1 increased for all detectors from 2018 to 2022. Among the top-performing detectors, the Autoencoder achieved the highest score gain – from 0.889 to 0.950 (an increase of 0.061). Isolation Forest gained from 0.814 to 0.882 (an increase of 0.068). LOF maintained a consistent good performance, increasing its score from 0.933 to 0.955 with a maximum of 0.960 in 2021. Detector scores increased across the simulated yearly cohorts. However, this trend may reflect the anomaly injection process. Increasing IOPS may also influence detection performance. Therefore, the results do not establish general detection improvements. Testing across different workload shifts is necessary. Further validation using genuine storage telemetry is also required. The performance of EWMA hardly changed from 0.616 to 0.631, meaning that the baseline smoothing is not enough when the baseline is changing. This upward trend is understandable due to the higher IOPS values in more recent years increasing the difference between the normal load and injected faults and making the anomalies more

visible. Moreover, the learned models also benefit from having a higher absolute deviation. This is confirmed by several reviews of anomaly detection literature suggesting that deep and unsupervised models are good with non-stationary data (Pang et al., 2021).

Statistical Test Results

Table 4. Omnibus Tests of Difference Among the Six Detectors

Test	Outcome variable	Statistic	df	p-value	Effect size
Friedman	F1	$\chi^2 = 219.13$	5	< .001	Kendall's W = 0.877
Friedman	AUC	$\chi^2 = 227.89$	5	< .001	W = 0.912
One-way ANOVA	F1	F = 768.64	5, 294	< .001	$\eta^2 = 0.929$

Table 4 indicates that the detector selection exerted an extremely powerful effect on the results. Friedman's test applied to F1 yielded $\chi^2(5) = 219.13$ ($p < .001$) and Kendall's W of 0.877. The 50 runs require consistent, matched conditions across all detectors. Independent runs must use separate random seeds and generated datasets. Each detector's parameters must be tuned using training and validation data. Test data must remain excluded from model selection. Friedman testing requires matched observations across all detectors. ANOVA assumptions must also be assessed before interpretation. These synthetic results cannot establish real-world storage performance. Validation using independent enterprise telemetry remains necessary. Friedman's test based on AUCs gave $\chi^2(5) = 227.89$ ($p < .001$), thus confirming the pattern for the rankings. ANOVA produced $F(5, 294) = 768.64$ ($p < .001$); hence, we have $\eta^2 = 0.929$ and 92.9% of variance in F1 explained by the method chosen. Shapiro-Wilk tests revealed some violations of normality (Isolation Forest, $p = .007$); accordingly, non-parametric Friedman's statistics is additionally justified alongside ANOVA. The null hypothesis of equal performances of the detectors is thus rejected with high confidence. Post hoc analysis is required in order to identify specific sources of the differences.

Post-Hoc Paired Comparisons Against Isolation Forest (Holm-Adjusted)

Comparator	Mean F1 difference (IF − comparator)	Wilcoxon W	Holm-adjusted p	Paired t	Cohen's d
Z-score	+0.015	341	.004	2.59	0.37
EWMA	+0.226	0	< .001	31.75	4.49
One-Class SVM	−0.089	0	< .001	−14.49	−2.05
LOF	−0.098	0	< .001	−16.64	−2.35
Autoencoder	−0.070	24	< .001	−9.47	−1.34

The Isolation Forest serves as the reference model in Table 5 due to its common application as a baseline. The Isolation Forest outperformed the EWMA method by 0.226 F1 score points ($d = 4.49$, $p < .001$), which is an enormous advantage. On the contrary, the difference in performance between the Isolation Forest and the Z-score method is very small ($d = 0.37$), however, it is statistically significant (Holm-adjusted $p = .004$). At the same time, Isolation Forest performed significantly worse than One-Class SVM ($d = -2.05$), Local Outlier Factor (LOF) ($d = -2.35$) and Autoencoder ($d = -1.34$) methods at $p < .001$ level. Holm correction has been made for 5 pairwise comparisons to control the family-wise error rate. The paired t-test and Wilcoxon test results coincided in all cases. Isolation Forest is computationally efficient, however, it has low recall on short anomalies.

McNemar Tests of Point-Level Classification Disagreement with Isolation Forest (n = 100,000 Test Points)

Comparator	IF correct, comparator wrong	Comparator correct, IF wrong	χ^2 (df = 1)	p-value
Z-score	1,914	1,584	30.94	< .001
EWMA	10,151	1,741	5,946.12	< .001

One-Class SVM	856	2,610	886.62	< .001
LOF	785	2,760	1,099.20	< .001
Autoencoder	1,407	2,789	454.52	< .001

Results for Table 6 have been generated based on classification for 50 runs using 2,000 test points per run, which leads to 100,000 classifications per detector and the number of instances when only one out of two detectors is accurate. The unique accuracy of Isolation Forest compared to EWMA is 10,151 versus 1,741 ($\chi^2 = 5{,}946.12$, $p < .001$), thus confirming a considerable advantage. Compared to Z-score the difference is 1,914 versus 1,584 ($\chi^2 = 30.94$, $p < .001$) in favor of Isolation Forest, a minor but significant advantage in terms of individual points. When compared to LOF the distribution is reversed with 785 versus 2,760 ($\chi^2 = 1{,}099.20$, $p < .001$). In turn, One-Class SVM and Autoencoder demonstrated the same tendency. Thus, we may conclude that the differences in F1 score are not an artifact of averaging. They stem from the differences in alerting decisions. In operations, such differences lead to missing incidents or false pages (Lu et al., 2020). The McNemar analysis uses 100,000 point-level classification decisions across repeated runs. However, observations within each run may be correlated. Therefore, treating every decision as independent may underestimate uncertainty. The independent sampling unit must be clarified. Future analysis should account for clustering across repeated runs. Confidence intervals should reflect run-level variation. Statistical significance requires cautious interpretation until this adjustment is completed.

Table 7. Year-Trend Tests on Isolation Forest F1 Across 2018–2022

Test	Statistic	df	p-value	Interpretation index
One-way ANOVA (Year)	$F = 4.71$	4, 45	.003	Means differ by year
Kruskal-Wallis (Year)	$H = 15.71$	4	.003	Medians differ by year
Spearman correlation	$\rho = .542$	$n = 50$	< .001	Moderate positive trend

Table 7 tests whether detection quality changed over time for Isolation Forest. The ANOVA, $F(4, 45) = 4.71$ ($p = .003$), and the Kruskal-Wallis test, $H(4) = 15.71$ ($p = .003$), both showed that F1 differed across the 5 yearly cohorts. The Spearman correlation of $\rho = .542$ ($p < .001$) indicates a moderate positive monotonic trend, consistent with the rise in mean F1 from 0.814 in 2018 to 0.882 in 2022. The five yearly cohorts represent simulated telemetry profiles, not production observations. Each cohort represents a different simulated workload condition. The 50 repeated runs must be mapped across these five cohorts. If ten runs were conducted per cohort, this allocation must be stated. Otherwise, the actual run distribution must be reported. The year labels indicate simulation categories, not observed calendar-year data. The agreement between parametric and rank-based tests strengthens confidence in this conclusion, given the non-normality noted earlier. The practical meaning is that anomaly detection became easier as workloads intensified, because injected faults deviated further from normal behavior. Real deployments may behave differently, so models should be retrained periodically to follow baseline drift, a point also emphasized in disk-failure prediction research spanning similar years (Lu et al., 2020; Xu et al., 2018).

Discussion

From the results obtained from the experiment, it is clear that detector selection plays a critical role in the quality of anomaly detection for enterprise storage telemetry. The selection of the detector accounts for 92.9% of F1 variance ($\eta^2 = 0.929$), and LOF (F1 = 0.948), One-Class SVM (0.939) and Autoencoder (0.920) detectors clearly performed much better than Isolation Forest (0.850), Z-score (0.835) and EWMA (0.624). It is in line with the findings that unsupervised and deep learning models perform better on multivariate drifting data (Chandola et al., 2009; Pang et al., 2021). Detection delay and daily alert volume are important operational measures. Detection delay measures the time between anomaly onset and detection. Daily alert volume measures alerts generated within each operational day. These metrics should accompany precision, recall, and F1-score. Their omission limits assessment of practical monitoring effectiveness. Future experiments should report both measures using timestamped predictions. The problem with the Isolation Forest was in recall (0.798). As anomalies within this

benchmark last from 1 to 5 points, are based on correlated metrics, the global partitioning of the dataset by this algorithm did not allow isolating the anomalies well enough (Liu et al., 2008). EWMA yielded precision 0.502, which means that approximately 49.8% of its alerts are false positives. All detector F1s showed an increase between 2018 and 2022, and Isolation Forest performance increased from 0.814 to 0.882 ($\rho = .542$, $p < .001$), due to the larger workload and thus better separation of normal behavior and fault patterns. However, this does not imply that detection gets easier in production systems. Research on proactive failure prediction suggests that production data is noisier and more unstable than simulations (Lu et al., 2020; Xu et al., 2018). There are 3 major limitations of this research. Synthetic nature of data, limitation of anomaly patterns to just 3 types, and setting of thresholds based on clean training set is uncommon practice in production environments. Strong statistical significance is attributed to the simulated nature of the experiment and should not be extrapolated before being validated. Operationally, density-based models or kernel approaches are advised for highly sensitive monitoring while Z-score serves as inexpensive preliminary filter. The future research direction include validation on real storage logs, concept-drift retraining schedule, and detection delay assessment in addition to F1.

Conclusion

Automated performance monitoring and anomaly detection could enhance the reliability and efficiency of enterprise storage. Existing methods based on threshold could ignore complex patterns in dynamic workloads. The current research seeks to compare Isolation Forest and autoencoder algorithms with traditional threshold-based monitoring. This study presents simulated experimental analysis of anomaly detection methods. Survey data collection remains planned future work. Therefore, no combined experimental and survey findings are claimed. Any future survey will require informed participant consent. Responses will remain anonymous and confidential. Participant data will be securely stored with restricted access. For the purpose of the experiments, the Microsoft Cloud Monitoring Dataset used along with metrics such as precision, recall, F1 score, false positive rate, and detection delay. An additional survey assess the experience and challenges of professionals in the area of monitoring. Thus, combining experimental data with survey results help obtain a more comprehensive insight into monitoring. It is important to note that cloud telemetry data might not reflect the specifics of dedicated enterprise storage. This limitation should be taken into account when interpreting the results of the research. Overall, the objective of the research is to select appropriate detection techniques for early warning, minimal number of false positives, and efficient resource management.

Reference list

1. Assaf, R., Giurgiu, I., Pfefferle, J., Monney, S., Pozidis, H., & Schumann, A. (2020). An anomaly detection and explainability framework using convolutional autoencoders for data storage systems. Proceedings of the Twenty Ninth International Joint Conference on Artificial Intelligence, 5228–5230. https://doi.org/10.24963/ijcai.2020/752
2. Assaf, R., Giurgiu, I., Pfefferle, J., Monney, S., Pozidis, H., & Schumann, A. (2020). An anomaly detection and explainability framework using convolutional autoencoders for data storage systems. Proceedings of the Twenty Ninth International Joint Conference on Artificial Intelligence, 5228–5230. https://doi.org/10.24963/ijcai.2020/752
3. Blázquez García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier and anomaly detection in time series data. ACM Computing Surveys, 54(3), Article 56. https://doi.org/10.1145/3444690
4. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier and anomaly detection in time series data. ACM Computing Surveys, 54(3), Article 56. https://doi.org/10.1145/3444690
5. Cao, C., Blaise, A., Verwer, S., & Rebecchi, F. (2022). Learning state machines to monitor and detect anomalies on a Kubernetes cluster. Proceedings of the 17th International Conference on Availability, Reliability and Security. https://doi.org/10.1145/3538969.3543810
6. Cao, C., Blaise, A., Verwer, S., & Rebecchi, F. (2022). Learning state machines to monitor and detect anomalies on a Kubernetes cluster. Proceedings of the 17th International Conference on Availability, Reliability and Security. https://doi.org/10.1145/3538969.3543810

7. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), Article 15. https://doi.org/10.1145/1541880.1541882
8. d Silva, M., Daniel, S., Kumarapeli, M., Mahadura, S., Rupasinghe, L., & Liyanapathirana, C. (2022). Anomaly detection in microservice systems using autoencoders. 2022 4th International Conference on Advancements in Computing, 488–493. https://doi.org/10.1109/ICAC57685.2022.10025259
9. Hushchyn, M., Sapronov, A., & Ustyuzhanin, A. (2019). Machine learning algorithms for automatic anomalies detection in data storage systems operation. Advances in Systems Science and Applications, 19(2), 23–32. https://doi.org/10.25728/assa.2019.19.2.725
10. Hushchyn, M., Sapronov, A., & Ustyuzhanin, A. (2019). Machine learning algorithms for automatic anomalies detection in data storage systems operation. Advances in Systems Science and Applications, 19(2), 23–32. https://doi.org/10.25728/assa.2019.19.2.725
11. Islam, M. S., Pourmajidi, W., Zhang, L., Steinbacher, J., Erwin, T., & Miranskyy, A. (2021). Anomaly detection in a large scale cloud platform. In 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE SEIP). IEEE/ACM.
12. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of the 2008 IEEE International Conference on Data Mining (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17
13. Lu, S., Luo, B., Patel, T., Yao, Y., Tiwari, D., & Shi, W. (2020). Making disk failure predictions SMARTer! In 18th USENIX Conference on File and Storage Technologies (FAST '20) (pp. 151–167). USENIX Association. https://www.usenix.org/conference/fast20/presentation/lu
14. Lu, S., Luo, B., Patel, T., Yao, Y., Tiwari, D., & Shi, W. (2020). Making disk failure predictions SMARTer! In 18th USENIX Conference on File and Storage Technologies (FAST ‘20) (pp. 151–167). USENIX Association. https://www.usenix.org/conference/fast20/presentation/lu
15. Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys, 54(2), Article 38. https://doi.org/10.1145/3439950
16. Xu, Y., Sui, K., Yao, R., Zhang, H., Lin, Q., Dang, Y., Li, P., Jiang, K., Zhang, W., Lou, J.-G., Chintalapati, M., & Zhang, D. (2018). Improving service availability of cloud systems by predicting disk error. In 2018 USENIX Annual Technical Conference (USENIX ATC '18) (pp. 481–494). USENIX Association. https://www.usenix.org/conference/atc18/presentation/xu-yong
17. Xu, Y., Sui, K., Yao, R., Zhang, H., Lin, Q., Dang, Y., Li, P., Jiang, K., Zhang, W., Lou, J.-G., Chintalapati, M., & Zhang, D. (2018). Improving service availability of cloud systems by predicting disk error. In 2018 USENIX Annual Technical Conference (USENIX ATC ‘18) (pp. 481–494). USENIX Association. https://www.usenix.org/conference/atc18/presentation/xu-yong