

dynamicMF: A Matrix Factorization Approach to Monitor Resource Usage in High Performance Computing Systems

Niyazi Sorkunlu
Computer Science and Engineering
University at Buffalo,
State University of New York
Buffalo, New York 14260
niyaziso@buffalo.edu

Duc Thanh Anh Luong
Computer Science and Engineering
University at Buffalo,
State University of New York
Buffalo, New York 14260
ducthanh@buffalo.edu

Varun Chandola
Computer Science and Engineering
University at Buffalo,
State University of New York
Buffalo, New York 14260
chandola@buffalo.edu

September 28, 2018

Abstract

High performance computing (HPC) facilities consist of a large number of interconnected computing units (or nodes) that execute highly complex scientific simulations to support scientific research. Monitoring such facilities, in real-time, is essential to ensure that the system operates at peak efficiency. Such systems are typically monitored using a variety of measurement and log data which capture the state of the various components within the system at regular intervals of time. As modern HPC systems grow in capacity and complexity, the data produced by current resource monitoring tools is at a scale that it is no longer feasible to be visually monitored by analysts. We propose a method that transforms the multi-dimensional output of resource monitoring tools to a low dimensional representation that facilitates the understanding of the behavior of a High Performance Computing (HPC) system. The proposed method automatically extracts the low-dimensional signal in the data which can be used to track the system effi-

ciency and identify performance anomalies. The method models the resource usage data as a three dimensional tensor (capturing resource usage of all compute nodes for difference resources over time). A dynamic matrix factorization algorithm, called *dynamicMF*, is proposed to extract a low-dimensional temporal signal for each node, which is subsequently fed into an anomaly detector. Results on resource usage data collected from the *Lonestar 4* system at the Texas Advanced Computing Center show that the identified anomalies are correlated with actual anomalous events reported in the system log messages.

1 Introduction

High performance computing (HPC) is at the forefront of scientific discovery and engineering innovation. To support the growing computing demands in these areas, modern HPC systems have seen a rapid transformation in terms of compute capacity and the underlying complexity. To ensure that an HPC system is operating at its peak effi-

ciency, analysts typically rely on system monitoring tools to monitor the system performance and identify anomalies.

Several tools exist for collecting and visualizing resource usage data from large scale HPC installations (e.g. Texas Advanced Computing Center *TACCStats* [10], XSEDE Metrics on Demand or XDMoD [17], etc.). Such tools can produce large amounts of high dimensional resource usage data at a high temporal frequency for each computational node in the system. The data collected by such tools provides a real-time view of the performance of the system which is typically fed into a visual interface (e.g., XDMoD [17], OVIS [3], Bright Cluster [4], etc.). However, such data is typically large and unwieldy, and visual monitoring is often challenging and inefficient.

Existing automated methods typically monitor each system resource or state for every compute node independently for potential deviations or anomalies [18], relying on a pre-defined threshold. Such methods can identify only those anomalous scenarios in which an individual node exhibits significant deviation for individual resources. These methods often miss anomalies that are under the threshold. Such anomalies, are weakly manifested across several nodes and multiple system resources and can be potentially detected by understanding the interactions between the different aspects of the system. A tensor decomposition based method [21] jointly models the interactions across the nodes, resources, and time, and produces a single time varying signal that can be used for tracking the overall system performance. However, the output does not allow for finer resolution analysis, for instance, identifying the specific nodes that are performing anomalously.

In this paper, we propose a method for understanding the system behavior at the node resolution. The key assumption here is that the observed system behavior, captured as resource usage information by tools such as *TACCStats*, can be decoupled into node and metric specific behaviors. Further, the node behavior can be decoupled into time-invariant (or static) and time-dependent (or dynamic) terms. This “decoupling” is achieved using a dynamic matrix factorization method which operates on the sequence of *node-usage* matrices, collected over time (See Figure 1). The output of the proposed algorithm is a set of low-dimensional representations for the metrics and the nodes that facilitate understanding of the system in

multiple ways. In particular, the dynamic node representation allows for tracking the performance of each node. Additionally, we use the proposed algorithm to produce a node level anomaly statistic and show that many performance anomalies, identified from message logs, co-occur with the identified anomalies.

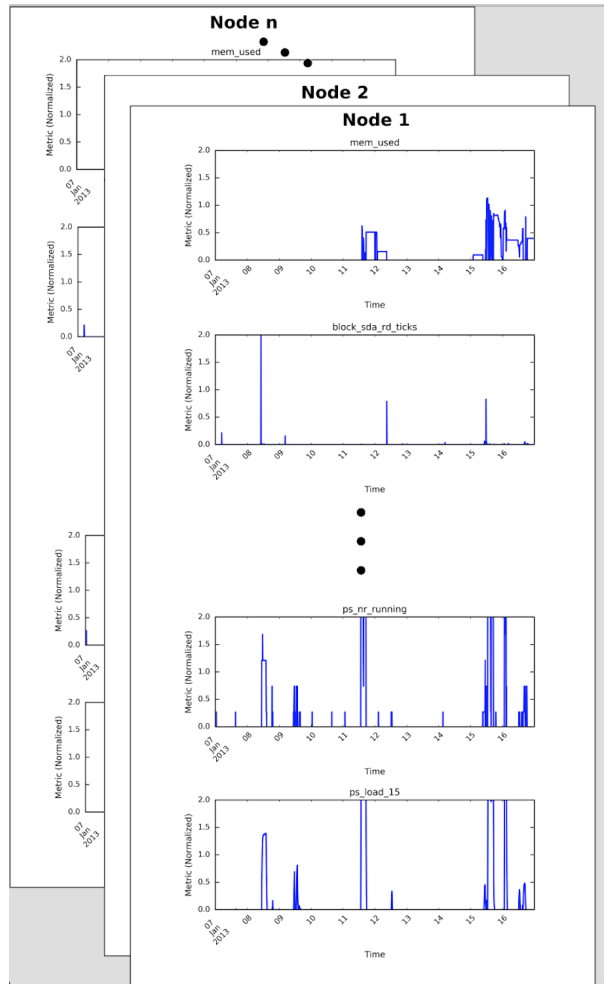


Figure 1: Typical resource usage data collected by tools such as *TACCStats*. Each panel corresponds to a compute node and consists of time-varying usage metrics for a variety of system resources.

The paper is organized as follows. The key problem addressed in this paper is outlined in Section 2, along with

the proposed model, *dynamicMF*. Results on a week long *TACC_Stats* data for the Lonestar 4 system in the Texas Advanced Computing Center are discussed in Section 3. Section 4 provides a brief overview of related works in this area and conclusion is presented in Section 5.

2 Proposed Methodology

We describe the proposed *dynamicMF* algorithm in this section. The algorithm is motivated by the fact that, while the performance of an HPC system is captured for many nodes and in the context of several resources, the metrics and the nodes exhibit a clustering pattern: several resources produce correlated usage metrics and many nodes behave similarly, either due to their computing specifications or the workload.

2.1 Problem Setting

We consider an HPC system consisting of N compute nodes. A resource usage monitoring program, such as *TACC_Stats* periodically reports usage data corresponding to M different resource usage metrics (CPU consumption, memory usage, network usage are three typical examples). At a given time t , the variable z_{nmt} denotes the m^{th} resource's usage for node n in the time interval ending at time t . We collect all of these variables at time t in a matrix $\mathbf{Z}_t \in \mathbb{R}^{N \times M}$, such that $Z_t[n, m] = z_{nmt}$.

2.2 Proposed Model

The proposed model assumes that the nodes and the metrics can be represented as *vectors* in two distinct (and unobserved/latent) K -dimensional spaces ($K \ll N, M$), respectively. Thus each metric is represented as a K -dimensional vector, $\mathbf{v}_m \in \mathbb{R}^K$, where each of the K dimensions denote a canonical characteristic of the corresponding metric. For the nodes, we assume two representations. The first representation is static over time, denoted as $\bar{\mathbf{u}}_n \in \mathbb{R}^K$, and characterizes the base behavior of the node. The second representation is time-dependent, denoted as $\hat{\mathbf{u}}_{nt} \in \mathbb{R}^K$, and captures the dynamic behavior of the node.

Assuming that the above representations, including $\{\mathbf{v}_m\}_{m=1}^M$, $\{\bar{\mathbf{u}}_n\}_{n=1}^N$, $\{\hat{\mathbf{u}}_{nt}\}_{n=1, t=1}^{N, T}$, are known, the ob-

served behavior for each node and metric combination, i.e., z_{nmt} , is generated by a functional interaction between the three representations. In this paper, we assume a linear interaction of the following form:

$$z_{nmt} = (\bar{\mathbf{u}}_n \odot \hat{\mathbf{u}}_{nt})^\top \mathbf{v}_m \quad (1)$$

where \odot is an element-wise product. For the entire resource usage matrix, the interaction can be written as:

$$\mathbf{Z}_t = (\bar{\mathbf{U}} \odot \hat{\mathbf{U}}_t) \mathbf{V}^\top \quad (2)$$

$\mathbf{V} \in \mathbb{R}^{M \times K}$ consists of the latent representations of the M metrics, i.e., $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M]^\top$. $\bar{\mathbf{U}} \in \mathbb{R}^{N \times K}$ consists of the static latent representations of the N nodes, i.e., $\bar{\mathbf{U}} = [\bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \dots, \bar{\mathbf{u}}_N]^\top$. $\hat{\mathbf{U}}_t \in \mathbb{R}^{N \times K}$ consists of the dynamic latent representations of the N nodes at time t , i.e., $\hat{\mathbf{U}}_t = [\hat{\mathbf{u}}_{1t}, \hat{\mathbf{u}}_{2t}, \dots, \hat{\mathbf{u}}_{Nt}]^\top$. Figure 2 illustrates the interaction between components of the proposed model to generate the observation for each metric and node pair in a given time window.

2.3 Inference Problem

The proposed generative model (See Figure 2) shows how the observations are generated through interactions between the three low-dimensional representations. However, given that these representations are unknown, the actual problem here is that of *inference*, i.e., given the observed resource usage data, $\{\mathbf{Z}_t\}_{t=1}^T$, estimate \mathbf{V} , $\bar{\mathbf{U}}$, and $\{\hat{\mathbf{U}}_t\}_{t=1}^T$.

Each of the inferred entities provides a distinct insight into the system performance. The latent metric characteristics (\mathbf{v}_m) map the metrics into a common space, where one can understand the similarity between the various metrics as inferred from the data. For instance, one would expect all metrics pertaining to the *network* resource appear similar in the latent space. However, how would *input/output* related metrics compare with the network metrics? Such relations are quantifiable using the latent space representation. Similar insights can be derived for the nodes using the static node characteristics ($\bar{\mathbf{u}}_n$). However, the dynamic node characteristics ($\hat{\mathbf{u}}_{nt}$) is the key output of the proposed model that allows for tracking the behavior of each node over time using a few latent characteristics. The information in matrices $\{\hat{\mathbf{U}}_t\}_{t=1}^T$ is obtained from the observed resource usage data after “explaining away” the

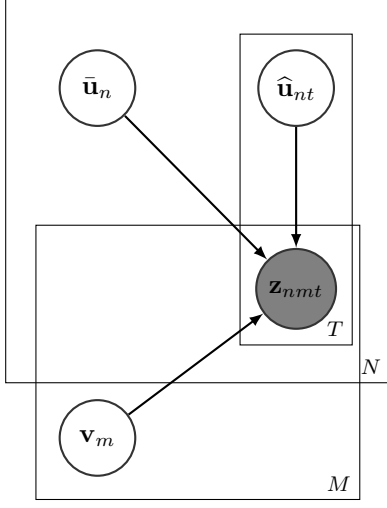


Figure 2: Illustration of the assumed interactions between the observed data (\mathbf{z}_{nmt}) and the latent variables ($\hat{\mathbf{u}}_{nt}$, $\bar{\mathbf{u}}_n$, \mathbf{v}_m) at a given time t for node n and metric m in the proposed *dynamicMF* model.

static node and metric behavior. Thus, the residual K -dimensional signal is expected to better provide information about the true dynamic behavior of each node.

2.4 Estimating Latent Representations

We pose the problem of estimating the latent entities, \mathbf{V} , $\bar{\mathbf{U}}$, and $\{\hat{\mathbf{U}}_t\}_{t=1}^T$, as an optimization problem that minimizes the error between the observed data ($\{\mathbf{Z}_t\}_{t=1}^T$) and reconstruction obtained in (2). The optimization problem can be written as:

$$\min \sum_{t=1}^T \left\| \mathbf{Z}_t - (\bar{\mathbf{U}} \odot \hat{\mathbf{U}}_t) \mathbf{V}^T \right\|_F^2 \quad (3)$$

with respect to $\bar{\mathbf{U}}, \mathbf{V}, \{\hat{\mathbf{U}}_t\}_{t=1}^T$

where $\|\cdot\|_F$ represents the Frobenius norm.

2.5 Optimization

In order to solve the optimization problem as stated in (3), we use Adam algorithm [15] as the main building block to

jointly optimize over the entire set of variables. The Adam algorithm performs first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments, and has been highly successful in the training of deep neural networks. Typically, as a gradient-based method, one step of Adam algorithm only requires one pass over the data. In addition, its optimization with momentum allows the algorithm to work well for mini-batches of data instead of the entire dataset. Beside its fast computation, this algorithm is also efficient in terms of memory usage and well-suited for training model with many parameters. Algorithm 1 provides the pseudo-code for our approach.

Algorithm 1 Adam Optimizer

Initialization: initialize the values of $\bar{\mathbf{U}}, \mathbf{V}$ and $\{\hat{\mathbf{U}}_t\}_{t=1}^T$
for $iter \in \{1, \dots, max_iter\}$ **do**
 Update $\bar{\mathbf{U}}, \mathbf{V}, \{\hat{\mathbf{U}}_t\}_{t=1}^T$ simultaneously using one step of Adam optimizer
end for

For the experiments conducted in this paper, we implement the *dynamicMF* algorithm using TensorFlow [1]. The hyper-parameters of the Adam optimizer are set by following the suggestion in the original paper [15], i.e. $\alpha = 0.001$ (step size), $\beta_1 = 0.9$, $\beta_2 = 0.999$ (decay rate in exponential moving average).

2.6 Anomaly Detection

By using the resulting latent values $\bar{\mathbf{U}}, \{\hat{\mathbf{U}}_t\}_{t=1}^T$ and \mathbf{V} , we can derive an anomaly score for node n at time step t which can be defined as a discrepancy between model and observation, i.e.:

$$\mathbf{a}_{nt} = \frac{1}{m} \sum_m |z_{nmt} - (\bar{\mathbf{u}}_n \odot \hat{\mathbf{u}}_{nt})^\top \mathbf{v}_m| \quad (4)$$

3 Experimental Results

In this section, we present the experimental results by applying the *dynamicMF* algorithm to analyze real resource usage logs for a large HPC system, and identify potential system anomalies.

3.1 Dataset

The dataset used in the experiments is a subset of data obtained from Lonestar 4 cluster located at the Texas Advanced Computing Center (TACC)¹. This cluster consists of 1888 computing nodes from which the records of 1709 nodes are used in this study. The resource usage data is collected using the *TACC Stats* system monitor [10] which records various resources usage statistics at each computational node every 10 minutes. In our experiments, we use a set of 86 resource usage statistics with a resolution of 10 minutes from 1:10:01 March 1st 2013 to 23:40:01 March 7th 2013. Totally, we have data of 1000 timesteps that will be used for analysis in our experiments. See table 1 for the complete list of resource usage statistics used in this paper. As different statistics have different scale and units, we have normalized them to have zero mean and unit variance.

In addition to resource usage statistics, we also have information of computing jobs that performed in Lonestar 4. In particular, for each computing node that the job was assigned to, the system records its start time and end time. There is also the set of system log messages that outputs log traces from programs running on the HPC system, including the messages from the applications running on the nodes.

3.2 Optimization performance and effect of different K

The performance of the optimization algorithm (See Algorithm 1) in estimating the latent representations is shown in Figure 3. Results are shown for $K = 3, 5$, and 10. For each K , the objective function decays with each optimization step and eventually converges.

As expected, using a larger number of latent dimensions results in a better fit. However, larger K also results in a larger number of latent variables to monitor. In this paper, we choose $K = 10$ to report the subsequent findings. Note that the average error between the model and observations is approximately 0.57 for $K = 10$ (See Figure 3 inset).

¹<https://portal.tacc.utexas.edu/archives/lonestar4>

Component	Resource usage metrics
CPU	user, nice, system, idle, iowait, irq, softirq
I/O	rd_ios, rd_merges, rd_sectors, rd_ticks, wr_ios, wr_merges, wr_sectors, wr_ticks
Lustre /scratch, /work	read_bytes, write_bytes, dirty_pages_hits, dirty_pages_misses, ioctl, open, close, mmap, seek, fsync, setattr, truncate, getattr, statfs, alloc_inode, setxattr, getxattr, listxattr, inode_permission, readdir, lookup, link, unlink, rename
Lustre network	tx_msgs, rx_msgs, tx_bytes, rx_bytes,
Virtual Memory	pgpgin, pgpgout, pgallocc_normal, pgfree, pgactivate, pgdeactivate, pgfault, pgmajfault, pgrefill_normal, pgsteal_normal, pgscan_kswapd, pgscan_direct, pginodesteal, slabs_scanned, kswapd_steal, kswapd_inodesteal, pageoutrun, allocstall, pgrotated

Table 1: List of resource usage metrics used for experiments.

3.3 Analyzing latent representations

In this section we illustrate how the latent representations obtained using *dynamicMF* can be used to gain insights about the underlying system. The first latent representation (\mathbf{V}) provides a characterization of the metrics. A two-dimensional visualization of the metric representation is shown in Figure 4. We use *Principal Component Analysis* (PCA) [14] to reduce the K -dimensional data to two dimensions for visualization.

Figure 4 reveals several interesting insights regarding the characteristics of various metrics. While metrics within most coarse categories appear to cluster together in the latent space (e.g., *llite*, *vm*), there are some clear outliers (e.g., *llite-work.listxattr*, *vm.pgmafault*).

In the same way, the static latent representations for

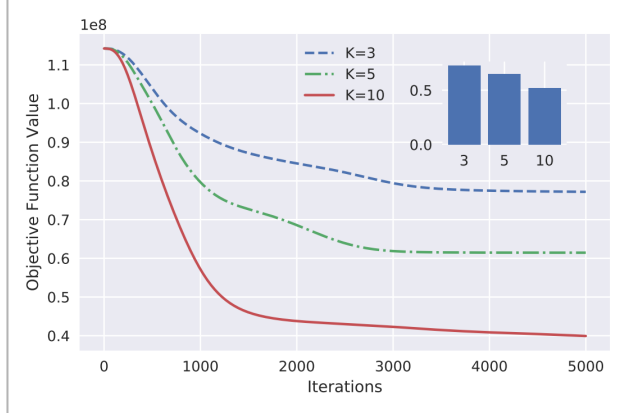


Figure 3: Objective function (See (3)) optimization for different number of latent dimensions (K). Inset shows the average final objective function value (or error) for z_{nmt} . (Best viewed in color.)

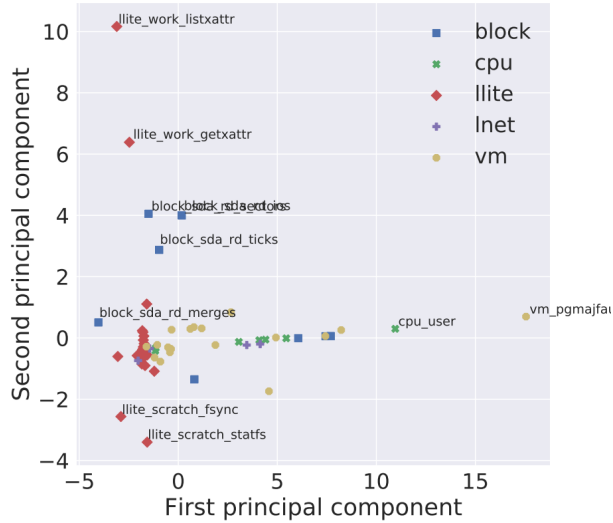


Figure 4: 2-dimensional view of the latent representation for metrics estimated by *dynamicMF* with $K = 10$. The K -dimensional data was reduced to two dimensions using PCA. Metrics belonging to a single coarse grouping (llite, vm, cpu, lnet, block_sda) are displayed in the same way. (Best viewed in color.)

the nodes in the matrix \bar{U} can be used to visualize the nodes. Figure 5 shows the two-dimensional plot for the static node behavior. An interesting pattern with two distinct and equal sized clusters is revealed. While a thorough post-analysis of the node specifications is needed to explain the patterns, we distinguish between the regular compute nodes with 24 GB memory and the large memory nodes with 1 TB memory. The large memory nodes are mapped in the periphery of two clusters, indicating that *dynamicMF* is able to capture the specification related characteristics of the nodes.

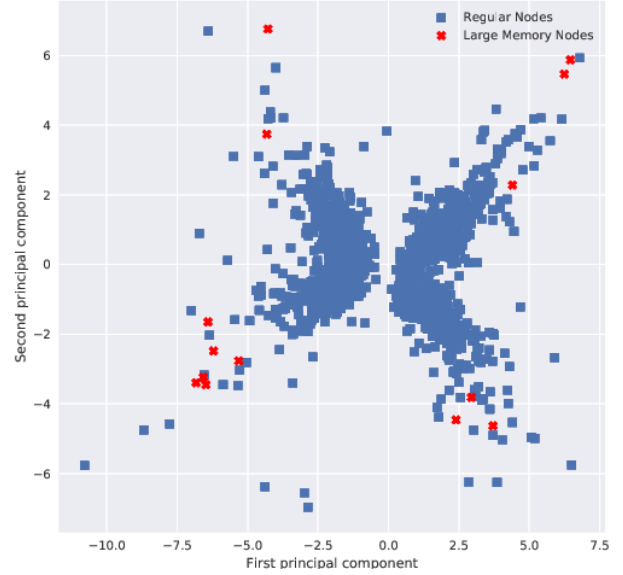


Figure 5: 2-dimensional view of the static latent representation for the compute nodes estimated by *dynamicMF* with $K = 10$, using PCA for reduction to two dimensions. Large memory nodes are displayed in a different way. (Best viewed in color.)

As mentioned earlier, the key output that is produced by *dynamicMF* is the dynamic latent representation for each node. An important point here is that each latent dimension contains unique information about the nodes. While we do not explicitly enforce orthogonality in the optimization formulation, the results show that the latent dimensions are not significantly correlated (See Figure 6).

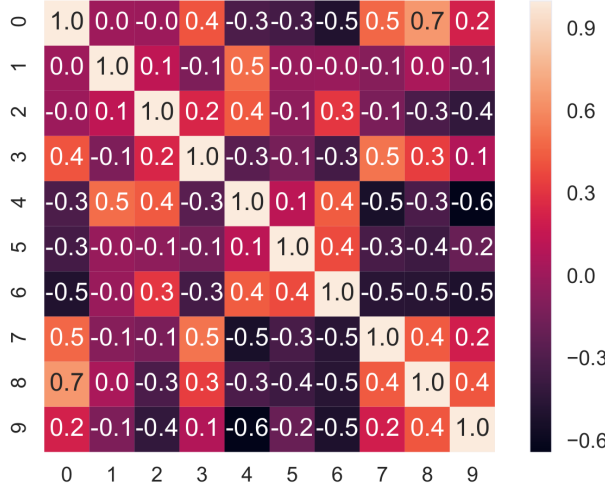


Figure 6: Pearson R correlation among the 10 latent node representations. (*Best viewed in color.*)

Figure 7 shows the K components of the time-varying \hat{u}_{nt} for two different nodes in the system. For both nodes, the different latent dimensions reveal a unique behavior of the node. Moreover, the output for the two nodes is significantly different. This is due to the fact that the two nodes had significantly different workloads during the target week (obtained from the job information present in the *TACC_Stats* data). Node `c325-312` (blue plot) had a significantly light workload with 9 executed jobs, and consequently the latent dynamic behavior has low variance. On the other hand, node `c336-203` (red plot) had a heavy workload with 60 executed jobs, which is reflected in the high variance for the latent dimensions. Figure 7 shows 10 latent dimensions of the time-varying component \hat{u}_{nt} for two different nodes with two different workloads in the system. In particular, node `c325-312` (blue plot) had a significantly light workload with seven jobs receiving from users. On the other extreme, node `c336-203` (red plot) had the most active jobs during the target week with 60 jobs in total. In Figure 7, these two nodes exhibit similar patterns of having sporadic spikes at some certain time steps. Although it looks counterintuitive to have similar patterns for nodes with two different workloads, one may note that even for node with zero job submitted from users in HPC system, it still has system-level jobs running

in the background which make the node have different resource usage at different time, causing sporadic spikes in the temporal behavior as seen in the figure. In addition, what we observe in Figure 7 is the latent temporal representation of nodes after “explaining away” the overall time-independent component \bar{u}_n . For this reason, spikes appear in the temporal representation as a result of unusual resource consumptions in each node.

3.4 Finding anomalies using *dynamicMF*

In this section, we analyze the association of our resulting hidden values and common errors obtained from system log messages. In particular, there are three frequent error types that we extracted from system log messages: (1) write error, (2) segmentation fault and (3) inode error. Write error is a type of error that happens in the system log with the format “*ost_write* operation failed with -122 ”. This is a linux error that happens when disk quota is exceeded². Segmentation fault is a common error type when a program tries to write or read in an invalid memory location. Inode error happens when *ll_inode_revalidate_fini()* fails and returns with error code -43 . We associate the error types from system log with our resource usage statistics by assigning the error that occurs within 10 minutes of the occurrence of the present time window.

In our analysis, we use the anomaly score defined in section 2.6 and analyze its correspondence to the frequent system log error types mentioned above. To compare, we also show results using an adapted anomaly detection method that constructs a three-way tensor using the resource usage data and then uses tensor reconstruction error as the anomaly score for the entire system [21]. While the original method provides a single score for the whole system, we adapt the method to produce a score for each node individually.

Figure 8 shows three typical examples of three computing nodes for each system log error type. In this figure, for write error, we do not observe any association between write error and anomaly score. As the write error happens as a result of exceeding disk quota, this kind of erratic behavior does not reflect in resource consumption and therefore it does not reveal any association with our

²<https://github.com/torvalds/linux/blob/master/include/uapi/asm-generic/errno.h>

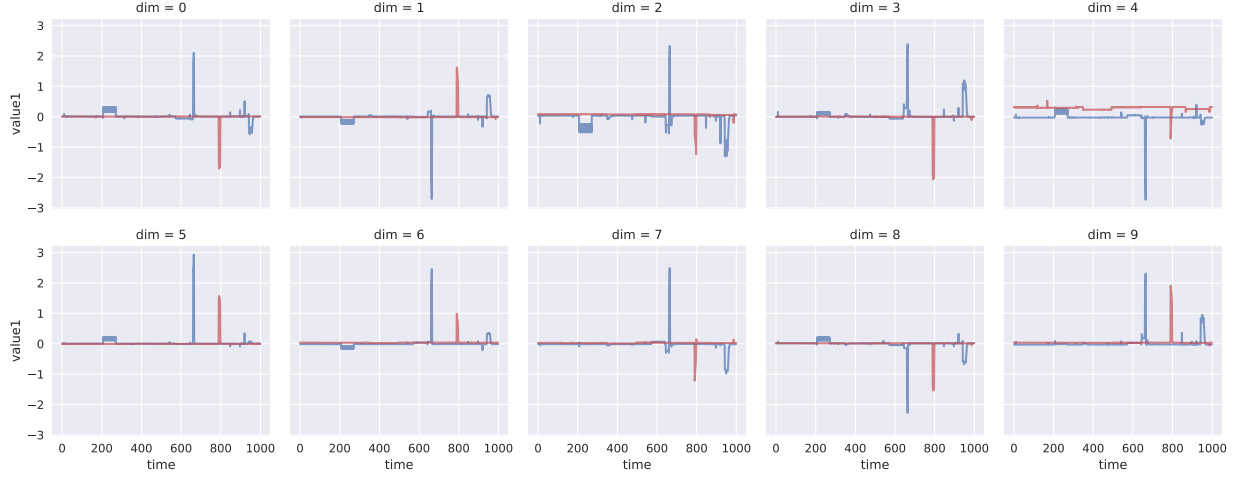


Figure 7: Temporal 10-dimensional representation estimated by *dynamicMF* for two selected nodes. Node c325-312 was inactive during the target week with seven executed jobs, shown in blue plot. Node c336-203 was the most active during the target week with 60 executed jobs, shown in red plot.. (Best viewed in color.)

residual values, as can be seen in Figure 8a. On the other hand, the segmentation faults in Figure 8b and inode errors in Figure 8c occur very close and even identical to the peaks of the anomaly score. As segmentation fault and inode errors are typical examples of buggy programs, these errors are usually results of faulty programming practice. This can lead to the programs with inefficient resource usage, which can be subsequently observed by our resource usage statistics. Therefore, in Figure 8b and 8c, the spikes in anomaly score values often appear roughly around the errors in terms of time.

The tensor based method (red plot in Figure 8) produces a metric shows high anomaly values that are temporally close to some of the errors. However, the method also produces a large number of false alarms, which can make such a method practically unusable.

4 Related Work

The dynamic matrix factorization algorithm proposed here falls under the general purview of *unsupervised representation learning* [2], that includes a variety of tasks such as dictionary learning, independent component analysis, autoencoders, matrix factorization and various forms

of clustering. In particular, a relevant topic is *sparse coding* [7], where the objective is to learn over-complete and sparse bases for a given data set. However, methods in these categories primarily focus on a static representation (typically as a matrix) and do not handle temporally evolving data.

The idea of dynamic matrix factorization has been used in designing the recommendation system. In particular, dynamic Poisson factorization (dPF) model works effectively when we assume each observation follows a Poisson distribution [6]. In this model, because of their probabilistic setting, variational inference has been used to perform inference. After that, they use the block coordinate approach to solve the optimization problem yield as a result of variational inference. Although our model is similar to dPF in the design of shared components and local components, our model is different from dPF in several ways: (1) we do not assume the temporal relationship between consecutive timestep as in dPF, (2) because of our non-probabilistic formulation, we can avoid the problem of non-conjugacy in dPF and use more effective gradient-based optimization such as Adam algorithm instead of using block coordinate optimization with L-BFGS (a more computationally intensive optimization

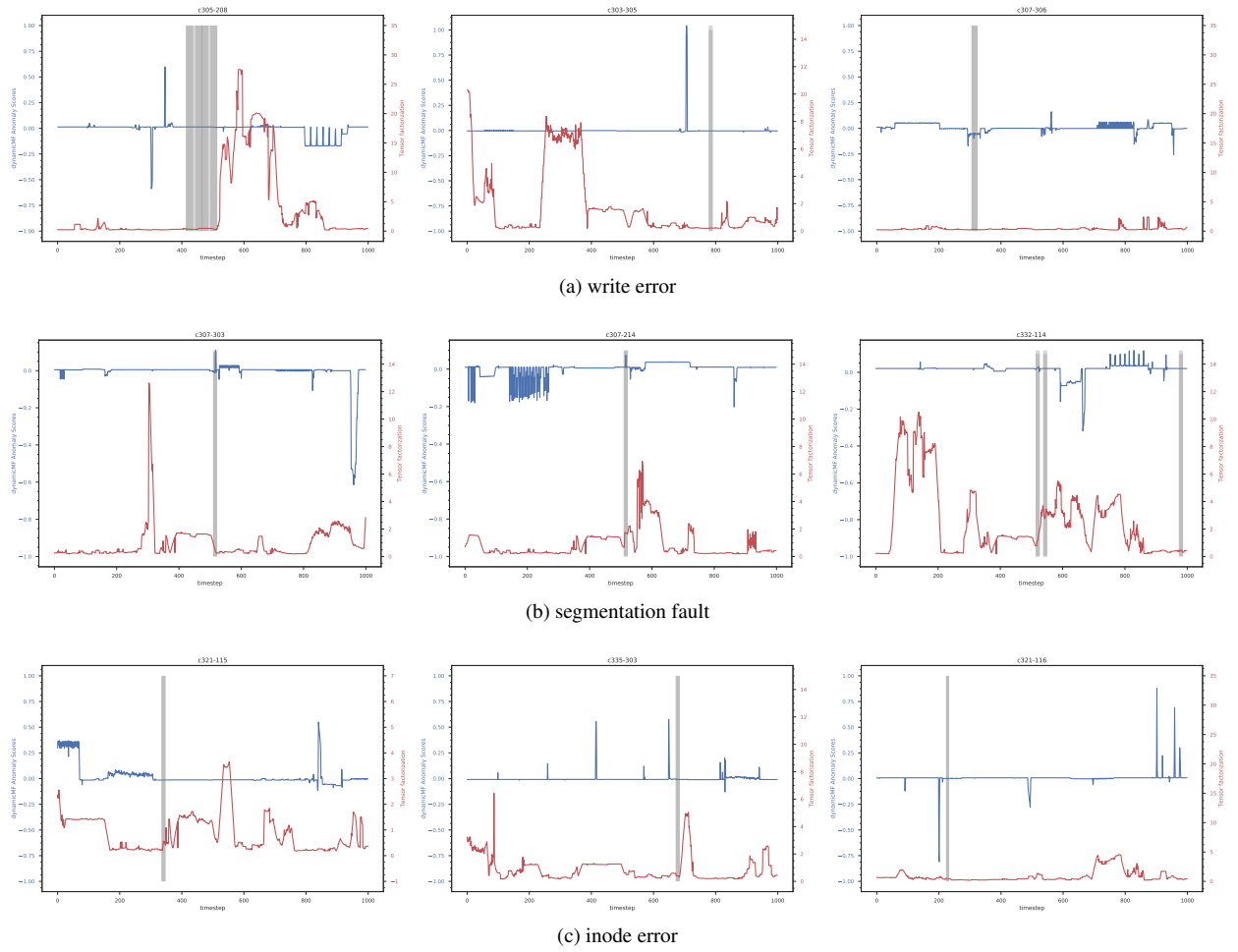


Figure 8: Association between anomaly scores computed using (4) and errors in system log messages. Anomaly scores produced by *dynamicMF* are shown in blue and the anomaly score produced by the state of art tensor reconstruction method is shown in red. Vertical gray lines indicate where the errors happen. The node name is shown above each figure. (*Best viewed in color.*)

method) as a building block.

Solutions for detecting, diagnosing, and predicting faults and failures in large high performance computing installations have typically relied on message logs [22, 16, 11, 20, 19] or resource usage data [12, 5], or both [9, 13, 8]. Since this paper focuses on the detection task, we present a brief overview of related methods that deal with detecting faults. Methods operating on message logs typically aggregate message logs by an entity of interest, e.g., a computational node/block [16, 22] or a job [11], and identify anomalous entities using different data representations, such as message arrival statistics [16], vector representation derived from the message content [22] or a state machine that models the dynamic behavior of the entity [11]. However, message logs are typically noisy and often incomplete, which has led to methods that analyze alternate data sources such as resource usage metrics or performance counter data [12, 5]. However, these solutions perform a node-specific or job-specific analysis of resource usage to identify anomalous nodes [12] or jobs [5]. Recently, solutions that combine resource usage data and message logs to improve fault detection have been proposed [13, 8]. The Crude system uses resource usage data to improve the performance of a PCA driven anomaly detection method [22] that operates at node and job level. However, none of the existing methods model the temporal dimension to better identify faults. In a recent work, a tensor based representation of the resource usage data was used to identify anomalies at the system level [21]. The method relies on low rank approximation of a three way tensor that captures interactions between nodes, metrics, and time, to establish an anomaly score for the entire system at every time instance. However, the method does not provide a node level estimate of system behavior and anomalies.

5 Conclusions and Future Directions

In this study, we propose *dynamicMF* - a dynamic matrix factorization method which takes both global and local behaviors of the system into account to analyze for variations. This method is applied to a list of time-varying resource usage matrices. In our experiments with real data

from Lonestar 4 system, *dynamicMF* helps reduce the dimension of resource usage statistics and provide visualization at a node-specific level. The proposed anomaly statistic allows for an easy identification of performance anomalies. Future direction of our research will be on developing a fully-automated anomaly detection method to detect failure in HPC system by using the low dimensional latent space obtained from *dynamicMF*.

Future direction of our research will be to leverage job information for analysis and incorporate it into our model. Successfully modeling the interaction between jobs and nodes can help us better explain the data and obtain a more accurate model to detect inefficiencies in resource usage.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.
- [3] J. Brandt, A. Gentile, J. Mayo, P. Pebay, D. Roe, D. Thompson, and M. Wong. Resource monitoring and management with ovis to enable hpc in cloud computing environments. In *2009 IEEE International Symposium on Parallel Distributed Processing*, pages 1–8, 2009.
- [4] https://www.microway.com/download/datasheet/Bright_Cluster_Manager_Brochure.pdf, 2017.
- [5] G. Bronevetsky, I. Laguna, B. R. de Supinski, and S. Bagchi. Automatic fault characterization via abnormality-enhanced classification. In *Proceedings of the 2012 42Nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 1–12, 2012.
- [6] L. Charlin, R. Ranganath, J. McInerney, and D. M. Blei. Dynamic poisson factorization. In *Proceed-*

- ings of the 9th ACM Conference on Recommender Systems, pages 155–162. ACM, 2015.
- [7] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.
 - [8] E. Chuah, A. Jhumka, J. C. Browne, N. Gurumdimma, S. Narasimhamurthy, and B. Barth. Using message logs and resource use data for cluster failure diagnosis. In *23rd annual IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC 2016)*, 2016.
 - [9] E. Chuah, A. Jhumka, S. Narasimhamurthy, J. Hammond, J. C. Browne, and B. Barth. Linking resource usage anomalies with system failures from cluster log data. In *Proceedings of the 2013 IEEE 32Nd International Symposium on Reliable Distributed Systems, SRDS '13*, pages 111–120, Washington, DC, USA, 2013. IEEE Computer Society.
 - [10] T. Evans, W. Barth, J. Browne, R. DeLeon, T. Furlani, S. Gallo, M. Jones, and A. Patra. Comprehensive resource use monitoring for HPC systems with TACC stats. In *International Workshop on HPC User Support Tools*, pages 13–21, 2014.
 - [11] Q. Fu, J.-G. Lou, Y. Wang, and J. Li. Execution anomaly detection in distributed systems through unstructured log analysis. In *International conference on Data Mining*, 2009.
 - [12] Q. Guan, D. Smith, and S. Fu. Anomaly detection in large-scale coalition clusters for dependability assurance. In *2010 International Conference on High Performance Computing*, pages 1–10, 2010.
 - [13] N. Gurumdimma, A. Jhumka, M. Liakata, E. Chuah, and J. Browne. Crude: Combining resource usage data and error logs for accurate error detection in large-scale distributed systems. In *2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*, pages 51–60, 2016.
 - [14] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.
 - [15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
 - [16] A. J. Oliner, A. Aiken, and J. Stearley. Alert detection in system logs. In *2008 Eighth IEEE International Conference on Data Mining*, pages 959–964, Dec 2008.
 - [17] J. Palmer, S. Gallo, T. Furlani, M. Jones, R. DeLeon, J. White, N. Simakov, A. Patra, J. Sperhac, T. Yearke, R. Rathsam, M. Innus, C. Cornelius, J. Browne, W. Barth, and R. Evans. Open XDMoD: A tool for the comprehensive management of high-performance computing resources. *Computing in Science Engineering*, 17(4):52–62, 2015.
 - [18] M. Peiris, J. H. Hill, J. Thelin, S. Bykov, G. Kliot, and C. Konig. Pad: Performance anomaly detection in multi-server distributed systems. In *Proceedings of the 2014 IEEE International Conference on Cloud Computing*, pages 769–776, 2014.
 - [19] A. Pelaez, A. Quiroz, J. C. Browne, E. Chuah, and M. Parashar. Online failure prediction for hpc resources using decentralized clustering. In *2014 21st International Conference on High Performance Computing (HiPC)*, pages 1–9, 2014.
 - [20] T. Reidemeister, M. A. Munawar, M. Jiang, and P. A. S. Ward. Diagnosis of recurrent faults using log files. In *Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, pages 12–23, 2009.
 - [21] N. Sorkunlu, V. Chandola, and A. K. Patra. Tracking system behavior from resource usage data. In *2017 IEEE International Conference on Cluster Computing, CLUSTER 2017, Honolulu, HI, USA, September 5-8, 2017*, pages 410–418, 2017.
 - [22] W. Xu, L. Huang, A. Fox, D. Patterson, and M. Jordan. Mining console logs for large-scale system problem detection. In *Proceedings of the Third Conference on Tackling Computer Systems Problems with Machine Learning Techniques*, SysML’08, pages 4–4, Berkeley, CA, USA, 2008. USENIX Association.