Forecasting for Cloud computing on-demand resources based on pattern matching

Eddy Caron, Frédéric Desprez, Adrian Muresan

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Forecasting for Cloud computing on-demand resources based on pattern matching

Eddy Caron, Frédéric Desprez and Adrian Muresan

N° 7217

February 2010

Domaine 3

ISSN 0249-6399 ISRN INRIA/RR--7217--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Forecasting for Cloud computing on-demand resources based on pattern matching

Eddy Caron, Frédéric Desprez and Adrian Muresan

Domaine : Réseaux, systèmes et services, calcul distribué
Équipes-Projets Avalon

Rapport de recherche n° 7217 — February 2010 — 24 pages

Abstract: The Cloud phenomenon brings along the cost-saving benefit of dynamic scaling. Knowledge in advance is necessary as the virtual resources that Cloud computing uses have a setup time that is not negligible. We propose a new approach to the problem of workload prediction based on identifying similar past occurrences to the current short-term workload history.

We present in detail the auto-scaling algorithm that uses the above approach as well as experimental results by using real-world data and an overall evaluation of this approach, its potential and usefulness.

Key-words: Cloud Computing, auto-scaling, pattern matching

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Prédiction des réservations de ressources de Cloud Computing à la demande par la méthode de reconnaissance de motifs

Résumé : Le Cloud Computing permet de bénéficier de l'extensibilité dynamique. Une connaissance anticipée des évènements est nécessaire afin de prendre en compte le temps non négligeable de mise en place des ressources virtuelles fournies par les plates-formes de Cloud Computing. Nous proposons une nouvelle approche à ce problème de prédiction de charge basée sur la corrélation d'événements passés similaires à l'historique à court terme de la charge observée.

Nous présentons en détail un algorithme de gestion automatique de l'extensibilité qui utilise cette nouvelle approche. Nous proposons également des expérimentations utilisant les traces de plates-formes réelles ainsi qu'une évaluation de cette approche.

Mots-clés : Cloud Computing, extensibilité automatique, reconnaissance de motifs

1 Introduction

The evolution of IT software services in the direction of Cloud Computing took a step forward in the efficient use of hardware resources through the use of virtualization. In a traditional hosting services the user receives a static amount of hardware resources that he or she makes use of. In contrast to this, the Cloud approach is to offer on-demand virtualized resources to its users. Because virtual resources can be added or removed at any time during the lifetime of the application hosted on a Cloud, the possibility of dynamic scaling arises. Even more, dynamic scaling can be easily automated either at Cloud provider level or ar Cloud client level through the use of the Cloud provider's APIs.

To take full advantage of the benefits of dynamic scaling, a Cloud client (user or middleware) needs to be able to make accurate decisions on when to scale up and down.

To achieve good performance, the Cloud client needs to be able to make accurate scaling decisions. These scaling decisions are influenced by several aspects as for example virtual resource setup time or migration of existing processes to free resources, but resource usage has the biggest impact on the decision.

The idea of self-similarity in web traffic is not new [4]. Based on this a new auto-scaling strategy can be elaborated. By identifying usage patterns that have occurred in the past and have a high similarity to the present usage pattern, a decision can be made as to the necessity and/or direction of scaling for the present situation.

This paper presents a new approach to the resource usage prediction problem based on identifying past patterns that are similar to the present use of the system. We present an algorithm for identifying the patterns by using an approximate matching approach.

In Figure 1 we have a generic Cloud system usage model to have a top-level view on the role of the prediction model. As part of a Cloud client's resource management module, the prediction module uses the Client's usage history to try and make an intelligent guess on short-term usage demands. This alone does not constitute the Client's scaling decision as there are a number of other relevant factors that should be taken into consideration like the migration of currently running tasks from virtual resources that need to be terminated. In the current work we are focusing only on resource usage prediction. The impact that other factors have on the scaling decisions of a Cloud Client is an interesting topic of research, yet it is beyond the scope of the current work.

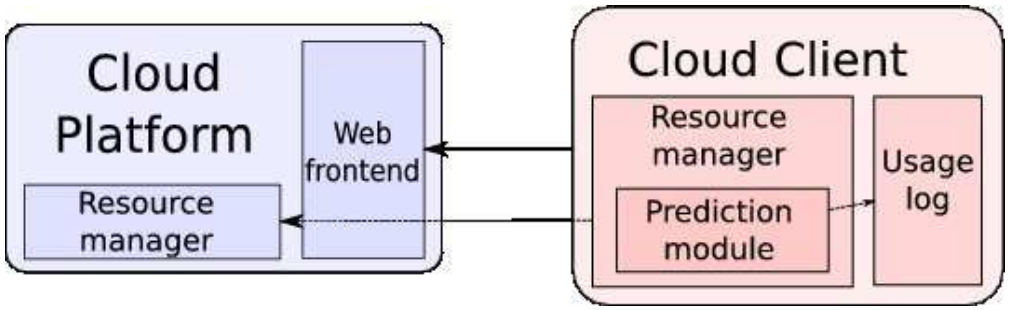

Figure 1: The role of the prediction component in a generic model of a Cloud system usage scenario.

The rest of this paper is organized as follows. The next section present an overview of existing approach given in the literature. Then, Section 3 presents our algorithm and its key design principles. Finally, before a conclusion and a description of future work, Section 4 presents our experimental results using actual grid traces.

2 Related Work

There are currently two main approaches for facilitating the auto-scaling decisions of Cloud client as a result of resource usage. The first approach treats the past server usage as a predictable sequence and constructs a mathematical model around it. As a result, the next value of the request sequence is obtained by evaluating the obtained model at the next time point. In other words, a prediction model is built by considering past

resource usage. The second approach is a reactive one, based on the current server load and auto-scaling rules that are set up by a human operator (usually a cloud client). This approach has been often referred to as the "Elasticity rules" approach or the "SLA" approach.

In [11] a description and comparison of three different auto-scaling algorithms is given: auto-regression of order 1 (AR1), Linear Regression, and the Rightscale algorithm. The auto-regression of order 1 algorithm is from the first category of auto-scaling algorithms. Its approach consists in using a finite history window and identifying appropriate parameters so that a recurring sequence can be obtained and therefore used to calculate the next values. The obtained parameters are adapted as the window slides along the time axis. The linear regression algorithm is also from the first category and calculates a polynomial approximation of the history of requests. The predicted value is then obtained by evaluating the polynomial at a higher point along the time axis. The Rightscale algorithm is from the second category, being a version of threshold-based auto-scaling. Its approach is to use a democratic voting system that is based on the current server load. Each virtual machine owned by the cloud client has a vote based on its current load level and two thresholds: low threshold that corresponds to a "scale down" vote (with a default value of 30% system usage) and a high threshold that corresponds to a "scale up" vote (with a default value of 85% system usage). The votes are collected by a central machine and the majority decides the scaling decision for the whole platform. The three algorithms have been put side-by-side and compared by a metric proposed in the same article. Their performance is considerably high.

A more complex form of SLA-based dynamic provisioning can be described by using elasticity rules that dictate what part of the cloud client needs to scale, in which direction and by how much. In [6], we find such an example with threshold-based rules. This is done by means of an extension to the OVF (Open Virtualization Format), an interoperable, platform and vendor neutral, open format that is used to describe VAs (Virtual Applications). VAs are preconfigured software stacks consisting of one or several Virtual Machines with the purpose of offering self-contained services. The OVF documnent is actually an XML document containing the description of the OVF package. The elasticity rules come as an extension of this document. They have three components: an associated name, a trigger condition based on the defined key performance indicators and an associated action that represents the concretization of the rule in the form of instantiating new components of the VA or removing existing component instances. Like the Rightscale algorithm, this approach is also a reactive one. Scalability rules have the benefits of combining the high performance of threshold-based algorithms such as Rightscale with tune-ability and therefore have been widely used in practice in commercial clouds.

In [1] a Decentralized Online Clustering model is described and proposed for automatic workload provisioning for enterprise grids and clouds and addresses their distributed nature. In this approach a workload prediction algorithm is used and integrated into the system to model the application dynamics. More specifically, a quadratic response surface model is used.

The ideas of workload prediction and workload modeling are by no means new, in fact they have been active areas or research in the field of Grid computing. In [10] we find a fine-detailed study on the topic of Grid performance evaluation by using synthetic workloads obtained from the modeling of grid workloads. The work describes performance metrics useful for evaluating grid environments. These are composed of traditional performance metrics that are time, resource or system related and grid-specific related to workload completion or failure metrics. The article continues by describing the specifics of grid workload modeling. These include user group modeling that underline the importance of taking into consideration statistics for all gobs on one hand and statistics for each user in particular on the other hand, based on his (or her) past actions. The article also describes submission patterns that arise in Grid environments and enumerate some of the current approaches of modeling them that include combining Poisson distributions for daily patterns or by using a polynomial function of degree eight. The authors argue that these pattern modeling approaches may not hold as they are indifferent to workload inter-dependency. The authors continue by presenting the GrenchMark [7] synthetic grid workload generation, execution and analysis framework. They also present extension sugestions to the framework that would make the framework be a better tool for workload generation and analysis.

In [9] we find an integration effort of a grid application development toolkit named Ibis [17], a grid co-scheduler name Koala [12] and the GrenchMark synthetic grid workload generator with the porpose of providing an end-to-end workload generation and testing framework. The authors argue about the benefits

that experimental testing of grid systems has over an analytical or simulated test model. The authors also argue in favor of using synthetic grid workloads over real grid workloads or benchmarking approaches. Next the authors describe their integration proposal of building applications with the Ibis toolkit, generating and submitting synthetic workloads with GrenchMark, and then scheduling them with Koala so that the results cal be analysed with GrenchMark again. As result of experimentation they concluded that workloads generated in GrenchMark can cover a wide range of run characteristics.

A non-linear model for grid workload prediction can be found in [5]. The authors propose a prediction model as a series of finite known functional components, usually taken from the sigmoig function class, with unknown coefficients. The coefficients are determined by using using the least square approximation method on a training set. The training set can be split into a training partition and an evaluation partition. This way an early stop strategy can be applied to avoid data overfitting. Their model has been tested on a 3D image rendering set of tasks based on the Blue Moon Rendering Tool. The error of their prediction is less than 14% with an average of 7.5%.

In [14] we find a real-time resource provisioning system for massive multiplayer online games based on a predictive usage model. The application is dynamically provisioned on a Grid environment. The authors propose a predictive model based on neural networks as this approach has more predictive power than simpler approches like exponential smoothing, yet is faster in terms of runtime than more complex approaches like autoregressive models, integrated models or moving average models. The neural network is prepared with two offline phases that include gathering of training samples and using them to train the neural network. As results of experimentation, the neural network approach has proven to have a greater accuracy when compared to the other tested prediction methods: average, moving average, last value and exponential smoothing. The obtained prediction error during the experiments has a maximum value of 33% and a minimum value of 4.94%.

3 String Matching based Scaling Algorithm

3.1 Idea Description

A Cloud client is provisioned depending on its use. The usage of a Cloud client can sometimes have a repetitive behavior. This can be caused by the similarities between tasks that the Cloud client is running or the repetitive nature of human behavior. Given the self-similar nature of web traffic it follows that current usage patterns of online services have a probability of having already occured in the past in a very similar form. Therefore we can infer what the system usage will be for a Cloud client by examining its past usage and extracting similar usages.

The pattern strategy has two inputs: a set of past Cloud client usage traces and the present usage pattern that consists of the last usage measures of the Cloud client. Cloud clients working in the same application domain have a higher similarity in resource usages. Due to this similarity it follows that the most relevant historic resource usage data that can be used comes from Cloud clients working in the same application domain. Therefore it would make sense to isolate historical data based on application domains before usage.

The present usage pattern of the Cloud client is used to identify a number of patterns in the historical set that are close to the present pattern itself. Identified patterns should not be dependent on their scale, just on the relation between the elements of the identified pattern and the pattern we are looking for. The resulting closest patterns will be interpolated by using a weighted interpolation (the found pattern that is closest to the present pattern will have a greater weight) and will have as result an approximation of the values that will follow after the present pattern. In essence, the usage of the Cloud client is predicted by finding similar usage patterns in the past or in other usage traces.

The problem of finding a pattern inside an array of data that is very similar to a given pattern is close to the problem of string matching. The approximate string matching problem has been widely studied especially in its relation to bioinformatics problems, yet it is considerably different from the problem we are addressing.

One definition for the approsimate string matching problem is the following: given a text string $T = t_0t_1...t_n$ and a pattern $P = p_0p_1...p_m$ find a substring of consecutive characters from T call it $T_i,_j$ that has the smallest edit (or Levenshtein) distance as possible [2].

1	A A	B B	A A	B B	A C	B	C
2	A	B A	A B	B A	A B	B C	C
3	A	B	A A	B B	A A	B B	C C

Table 1: KMP example

P =	A	B	A	B	A	C
π =	_	_	0	1	_	

Table 2: Calculating the auxiliary array

The edit distance is defined as the number of simple string operations: insert, delete, replace and sometimes exchange, that needs to be performed on the identified text substring to have equality to the pattern. The operations can have the same or different weights, depending on problem needs. The identified match can have any length because of the possible insert and delete operations.

In the problem that we are addressing, the edit distance cannot be applied as we are not comparing string character values, but floating point values. We are interested in identifying sub-arrays of the same over very close length and and whose floating point absolute value difference is as close as possible to zero. An insertion into or deletion from the identified sub-array would have a great impact on the floating-point difference.

We shall now describe the problem of string matching and its relation to the problem that the current paper addresses, as well as our proposal for the approximate variant that is relevant to our problem.

3.2 The String Matching Problem

The string matching problem consists in finding the position of a string (called pattern) inside a larger string. There are several approaches solutions to this well known problem. We have chosen the Knuth-Morris-Pratt (abbreviated KMP) as its performance are good as described in [3]. The KMP algorithm consists of a preprocessing step with a running time of $\Theta(m)$ where m is the length of the matching pattern and a matching step with running time of $\Theta(n)$ n where is the length of the string to match against. The algorithm is also embarrassingly parallel as it is data independent. Therefore the input data can easily be divided into independent blocks on which the algorithm can run in parallel.

The efficiency of the KMP algorithm is due to its approach in saving unnecessary comparisons in case of a mismatch between the pattern and the string to match against. It is able to do this by first identifying repetitive prefixes of the input pattern in the preprocessing step.

Consider the following example: input pattern P = “ABABC” and matching string T =“ABABABC”. There are three possible positions for P to be found in T, by using a sliding window approach, until one of the matches succeeds.

In the example given in Table 1, when step 1 fails, the pattern slides to the next possible position in the matching string and a new comparison is made in step 2. After step 2 fails, the pattern slides once again and reaches step 3 which makes a full match.

In Table 1, step 2 can be skipped altogether if we consider the relation that the pattern has with itself, i.e. its repetitive prefix. Once the first 4 characters of P have been matched against the 4 consecutive characters in T (the following 4 characters starting from position 0) we deduce that there is no need to restart the whole matching from position 1 in T because, from analyzing P we know that the match will fail as the 4 characters of T starting from 0 are the same as the first 4 characters of P starting from 0.

To assist the matching process, an auxiliary array is constructed over P (called π) that contains at position i, the ending position of the largest prefix of P that is a suffix of P[0..i]. For the P in our example, we have the results given in Table 2.

The entries in π that have a value of ”_” represent entries that are not prefixes of P. For example the second “A” in P is both a prefix and also a suffix of P[0..2] = “ABA”. The largest prefix that is also a suffix for P[0..2] is “A” and has the ending position at P[0]. This means that once we have matched P[0..2] = “ABA” in T and P[3] does not match in T, we can continue matching in T from the same index of T, and

we can start in P knowing that we have already matched the first character in P, as it is a prefix of length 1 of P[0..2]. Therefore we resume matching with the P index of $\pi[2] +1 = 0 +1 = 1$. Now, resuming our matching example, in step 1 we have matched P[0..3] to T[0..3]. We have P[4] != T[4], but we know that P[0..3] = "ABAB" has "AB" as the largest prefix that is also a suffix. So we can resume by matching T[4] to P[π[3]] = P[1], skipping P[0].

The preprocessing step The goal of the preprocessing step is to compute the π array. At each index i, π stores the end position of the longest prefix of P[0..i], that is also a suffix of P[0..i]. The algorithm for this has a runtime of $\Theta(m)$ where m is the length of P (Algorithm 1).

Algorithm 1 Calculate-prefix(P)

```
1: m ← length(P)
2: π[0] ← -1
3: k ← -1
4: for q ← 1 to m - 1 do
5:    while k > -1 and P[k+1] ≠ P[q] do
6:       k ← π[k]
7:    end while
8:    if P[k+1] = P[q] then
9:       k ← k+1
10:   end if
11:   π[q] ← k
12: end for
13: return π
```

The matching step The matching algorithm (Algorithm 2) has a runtime of $\Theta(n)$, where n is the length of T, the string to match against. It is very similar to a naive matching algorithm, but improved to skip redundant comparisons.

Algorithm 2 KMP(T, P)

```
1: n ← length(T)
2: m ← length(P)
3: π ← Calculate-prefix(P)
4: q ← -1
5: for i ← 0 to n - 1 do
6:    while q > -1 and P[q+1] ≠ P[i] do
7:       q ← π[q]
8:    end while
9:    if P[q+1] = T[i] then
10:      q ← q+1
11:      if q = m-1 then
12:         write "Found at position" i-m
13:         q ← π[q]
14:      end if
15:   end if
16: end for
```

3.3 Algorithm Description

The KMP Algorithm (2) is a good solution to the string matching problem. Despite the great similarities, our own pattern matching problem has some particularities of this own:

1. an approximate matching is needed since the odds of finding an identical pattern to the one we are looking for are considerably low;
2. matches which are too dissimilar either on small intervals or as a whole need to be discarded;
3. when comparing the pattern to the matching data, scale also needs to be taken into consideration. To be more exact, the scale of the pattern and the scale of the possible match should not affect the comparison, therefore it needs to be scale-independent.
4. the resulting matches are interpolated having different weights on the final result, based on their similarity to the identified pattern.

In order to do an approximate matching, the original KMP algorithm needs to be changed in the content of both functions, therefore they need to be modified accordingly.

Two types of approximation errors are used for the matching:

1. an instant error which dictates the amount by which the current match is allowed to differ from the pattern by comparing in smallest possible units;
2. a cumulative error that characterizes the amount by which the current match is allowed to differ from the pattern as a whole. This is basically a sum of the instant errors of the whole matching.

Figure 2 illustrates graphically the difference between the two types of acceptable errors (instant and cumulative) when comparing two patterns.

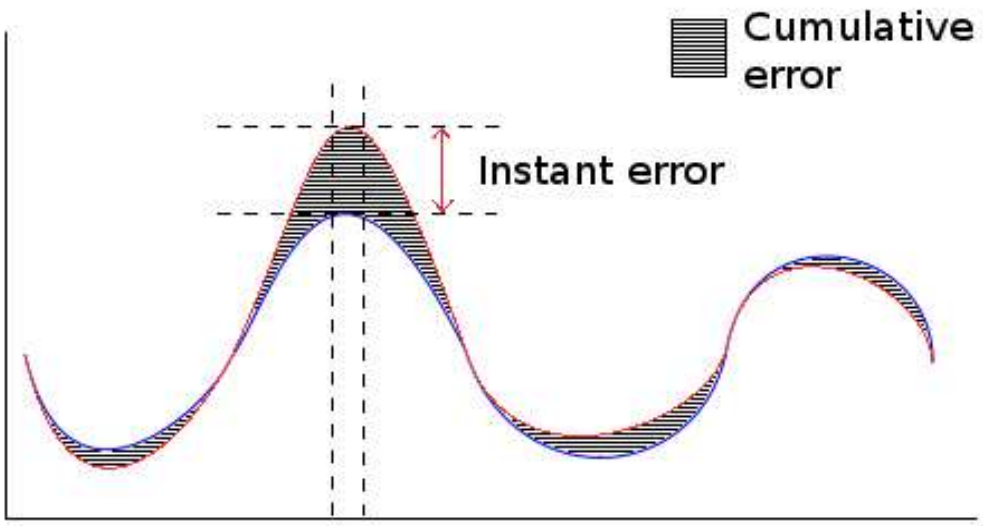

Figure 2: Difference between the two types of acceptable errors.

3.3.1 Scale-Independent Comparison

The distance between the pattern we are trying to match and a candidate pattern should be computed in a scale-independent manner by first normalizing the two pattern values to a common scale. To decrease floating point approximation errors, one can choose a distance computation that does not use divisions and therefore calculating only on integer values.

As an example, consider the pattern and the candidate from Figure 3. The pattern is an array containing values: 20 , 38 , 21 and the candidate match contains values: 42 , 81 , 39 . In this form, we cannot compare the two patterns. A first idea would be to normalize both arrays to a floating point [0..1] interval and then compare. Working with floating point numbers can be avoided by working with big integer numbers. To reach a common scale we simply multiply each array by the scale of the other. For the scale of each array

we can simply consider the first element. As a result, the pattern array is multiplied by the scale of the candidate (this is 42) and the candidate is multiplied by the scale of the pattern (which is 20). The result is depicted in Figure 4. In this new situation, comparing two components of each array is done simply by subtraction. The instant error is used here to assure that there are no two components that differ two much (in percentage) from the two arrays.

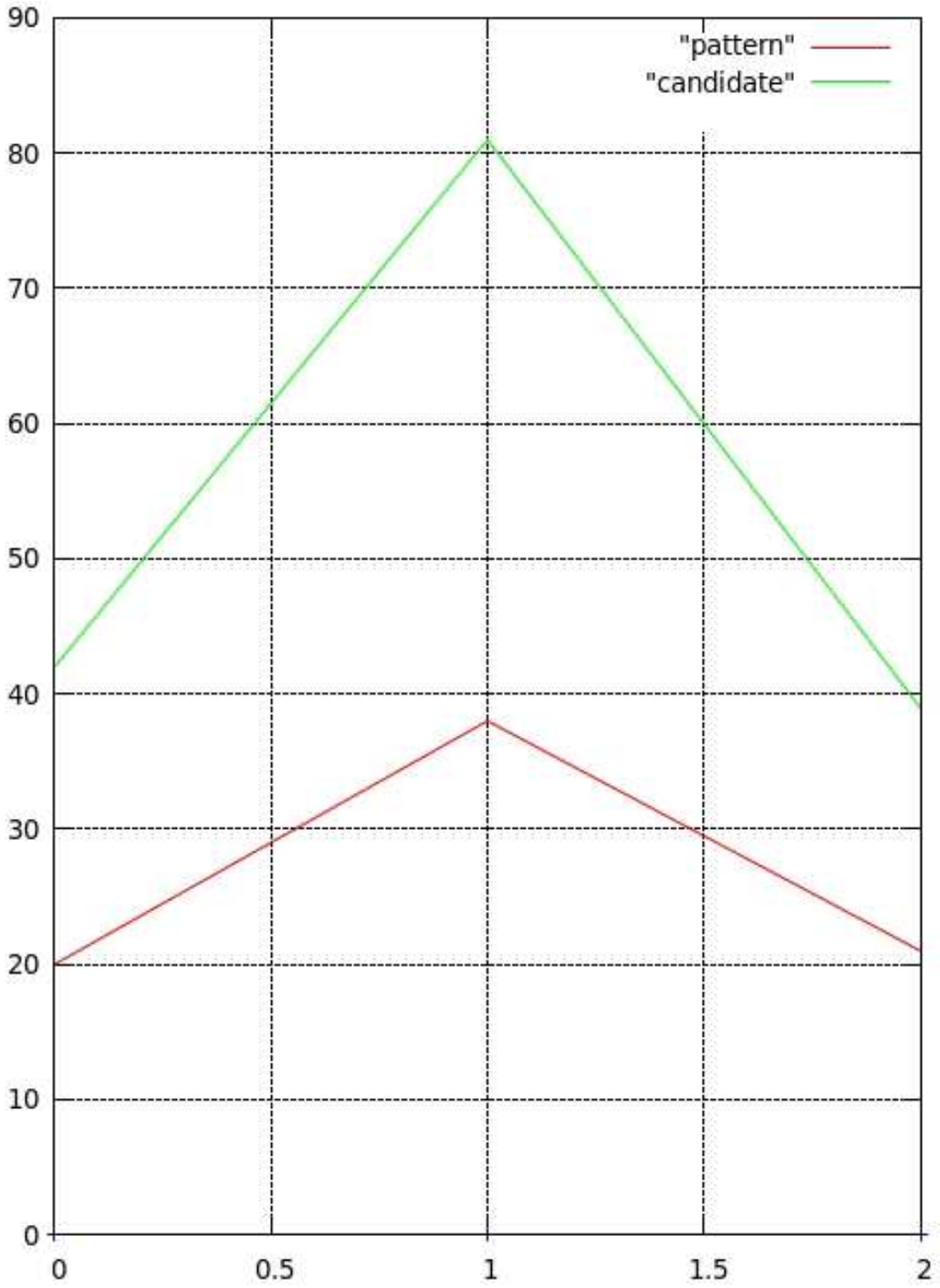

Figure 3: Scale-independent comparison - initial

Figure 4: Scale-independent comparison - common scale

Once the comparison is done, the identified candidate is stored along with its total distance from the pattern. This facilitates the significance of the result, as the candidate that is closest to the pattern has a higher weight in final result. The pseudocode for computing the instant error is illustrated Algorithm 3.1 in the Distance function.

Algorithm 3.1 Distance(PatternElement, PatternScale, DataElement, DataScale)

return
PatternElement × DataScale
- DataElement × PatternScale

The cumulative error is obtained by summing up the instant errors from all the elements of the pattern and candidate. This is illustrated in the CumulativeDistance function.

Algorithm 3 CumulativeDistance(P, T, DataOffset)

```
1: patternScale ← P[0]
2: dataScale ← T[DataOffset]
3: length ← length(P)
4: distance ← 0
5: for index ← 0 to length do
6:    distance ← distance + | dataScale × P[index] - patternScale × T[index + DataOffset] |
7: end for
8: return distance
```

3.3.2 KMP Modification

Algorithm 4 Calculate-prefix-approx(P, ACCEPT_INST_ERR)

```
1: m ← length(P)
2: π[0] ← -1
3: k ← -1
4: scaleK = P[0]
5: scaleQ = P[1]
6: for q ← 1 to m - 1 do
7:    dist ← Distance(P[k+1], scaleK, P[q], scaleQ)
8:    maxDistance ← ACCEPT_INST_ERR × scaleQ × P[k+1]
9:    while k > -1 and dist > maxDistance do
10:      k ← π[k]
11:      dist ← Distance(P[k+1], scaleK, P[q], scaleQ)
12:      scaleQ = P[q - (k+1)]
13:   end while
14:   if dist ≤ ACCEPT_INST_ERR × scaleQ × P[k+1] then
15:      k ← k+1
16:   end if
17:   π[q] ← k
18: end for
19: return π
```

The prefix calculation function is changed as described in Algorithm 4. The scales of the two components compared are represented by the first value of each component. This is arguable, but in practice we have achieved good results with this approach. In the function, scaleK represents the scale of the prefix and scaleQ represents the scale of the postfix of the pattern. The Distance function returns an appreciation of the distance between two different pattern instances, each having a different scale which is passed as parameter. The comparisons on lines 9 and 14 assure that the current instant distance does not differ by more then the acceptable error (in percentage) from the actual pattern that we are matching. The scaleQ

term, representing the scale of the data, from the comparison is needed for bringing the current term of the pattern to the same scale as the data.

The matching algorithm is changed as described in Algorithm 5. The main difference when compared to the original KMP algorithm is the use of the instant and cumulative distances as a means of filtering out potential matches that are too different either on small time intervals or as a whole.

Algorithm 5 KMP-approx(T, P, ACCEPT_INST_ERR, ACCEPT_CUMUL_ERR)

```
 1: n ← length(T)
 2: m ← length(P)
 3: π ← Calculate-prefix(P)
 4: q ← -1
 5: scaleP = P[0]
 6: scaleT = T[0]
 7: for i ← 0 to n - 1 do
 8:    dist ← Distance(P[q+1], scaleP, T[i], scaleT)
 9:    maxDist ← ACCEPT_INST_ERR × scaleT × P[q+1]
10:    while q > -1 and dist > maxDist do
11:       dist ← Distance(P[q+1], scaleP, T[i], scaleT)
12:       q ← π[q]
13:       scaleT = T[i - (q+1)]
14:       maxDist ← ACCEPT_INST_ERR × scaleT × P[q+1]
15:    end while
16:    if dist ≤ maxDist then
17:       q ← q+1
18:    end if
19:    if q = m-1 then
20:       dist ← CumulativeDistance(P, T, i - m + 1)
21:       maxDist ← ACCEPT_CUMUL_ERR × patternSum × scaleT
22:       if dist ≤ maxDist then
23:          StoreSolution(dist / scaleT, i - m + 1)
24:       end if
25:       q ← π[q]
26:       scaleP = P[q+1]
27:       scaleT = T[i - (q+1)]
28:    end if
29: end for
```

On lines 10 and 16 we ensure that the instant distance between the identified candidate and the pattern is no more than what the acceptable error permits. In order to ensure a correct comparison, the pattern term needs to be scaled to the same size as the data, hence the scaleT term is used in the comparison. Filtering by cumulative distance is done in lines 20 to 24. The CumulativeDistance function returns a sum of instant distances for every instant of the two compared arrays. The running time of this function is $\Theta(m)$ where m is the length of the arrays, which in our case is always equal to the length of P. Line 22 of the algorithm assures that the cumulative distance of the candidate does not differ more than is accepted by the cumulative error from the pattern itself. The pattern itself is represented by the patternSum term in the comparison. This is a sum of all the terms in the pattern and should be calculated only once, at the beginning of the algorithm. The pattern sum needs to be brought to the same scale as the candidate sequence and therefore the scaleT term is used. Filtering by an acceptable cumulative error that is smaller or equal to the acceptable instant error is useless. This conclusion is trivial when taking into consideration that the cumulative error is a sum of all the instant errors.

The use of the cumulative error changes the running time of the matching algorithm to $\Theta(n \times m)$ in the worst case, where n is the length of the string to match against and m is the length of the input pattern.

3.3.3 Interpolating the Values Found

Once approximate matches have been found, the problem of obtaining a relevant result from those matches is raised. Each match should have a contribution to the final result that is proportional to its relative distance to the pattern with respect to the other identified patterns. This corresponds to a weighted sum of the identified matches, where weights are calculated by considering the distance of the current match to the pattern and to the rest of the matches. Once the weights are calculated, the interpolation is performed between the following L elements after each approximate match. The result is a predicted sequence of length L.

3.3.4 Algorithm parameters

The algorithm accepts a number of parameters used for fine-tuning in accordance to each use-case. These parameters are:

- The maximum number of matches (called closest neighbors) to take into consideration (denoted K).
- The length of the predicted sequence (denoted L).
- The acceptable instant error representing the amount by which the identified sequence is allowed to differ on the smallest possible interval lengths from the pattern we are looking for.
- The acceptable cumulative error which represents the amount by which the identified sequence is allowed to differ as a whole from the pattern we are looking for.
- The input set of data representing the database of past requests.
- The input pattern representing a sequence with the last period of requests received.

The first parameter is not independent of the others. It is actually influenced considerably by the acceptable errors. The correlation is strong and can be expressed very easy: the larger the acceptable error, the more matches the algorithm identifies, but the more irrelevant they will be.

Calculating the acceptable errors

The value of the acceptable errors can be calculated based on the maximum number of neighbors that we wish to find. The approach for this is to use a binary search to zone in on the appropriate values for the acceptable errors. By using the binary search approach, we have obtained values that have proven to be good in practice. We have used a lower bound of 20% of K for a minimum of identified neighbors and 90% of K as the upper bound for maximum number of identified neighbors.

Calculating the appropriate pattern length

The length of the pattern that represents the last traces of server usage has a great impact on the results of the algorithm. Finding the appropriate length is a problem in its own as we have a trade-off between patterns of big lengths that yield a small number of similar candidates, that might be to small in order to be usable, and patterns of small lengths, that find more candidates but they tend to be more irrelevant to our current situation. We have chosen two approaches to this problem. The first approach is to find the most lengths of the most frequent repetitive patterns and use the same length as input to the prediction algorithm.

We have the following constructive approach to identifying the length of the most frequent repetitive patterns:

1. find all similar patterns of length 2 in the historic data
2. take all similar patterns of length 2 and try to match the next element too. This yields all similar patterns of length 3.

   ...
3. take all patterns of length n and try to match the next element too. This yields all similar patterns of length $n + 1$.

	LCG	**Nordugrid**	**SHARCNET**
Avg	8970	91893	33516
Min	0	0	0
5%Min	69	11	73
10%Min	79	24	152
Median	255	3861	12165
Max	586702	1452763	7449415
Tested	200 2500 5000 10000 50000 100000	10000	10000

Table 3: Job length statistics on different grid traces. Values represent time in seconds based on the running time of the recorded jobs.

The result is that the number of identified similar patterns decreases as the length of the patterns increases:

$$count[n+1] \leq count[n] \leq ... \leq count[3] \leq count[2]$$

The conclusion is that the most frequent patterns are of the ones with length 2. In practice, using a pattern length of 2 would have the following consequences:

- Good for predicting very short in advance (i.e. 1)
- Loses meaning when trying to predict longer sequences
- The idea of trend is lost as the steps are very small while the trend is a longer sequence.

We need to have a better way for choosing the pattern length, that would give more relevant results and avoid pollution as much as possible.

The length of the pattern should be influenced by the time it takes to service a request on the server. We then have the following possibilities:

- Median / average
  - Representative of most of the requests
- Minimum
  - A large pattern cannot match against a smaller pattern that is half different
  - A small pattern can match against a large pattern that's half different
  - The minimum is very probably close to 0 (grid testing experiments)
  - A close minimum can be selected (ex. 5% - 10% from the bottom)

By using real-world grid traces from the workload archive of TUDelft University [16]. We have used the running time in seconds of each job and obtained the results given in table 3. We have tested traces from several research grids [8, 13, 15] to get a real-world appreciation of possible values for the pattern length, by taking different metrics.

We can also consider plots of sorted job lengths in seconds.

The conclusions given by Figures 5, 6, and 7 along with the previous table are that, for all practical purposes, a pattern length that is a minimum or even median of the time it takes to service a request, is unusable when dealing with servers that have a similar usage to the research grids described above. In practice we have used the average of the request service time and have obtained good results.

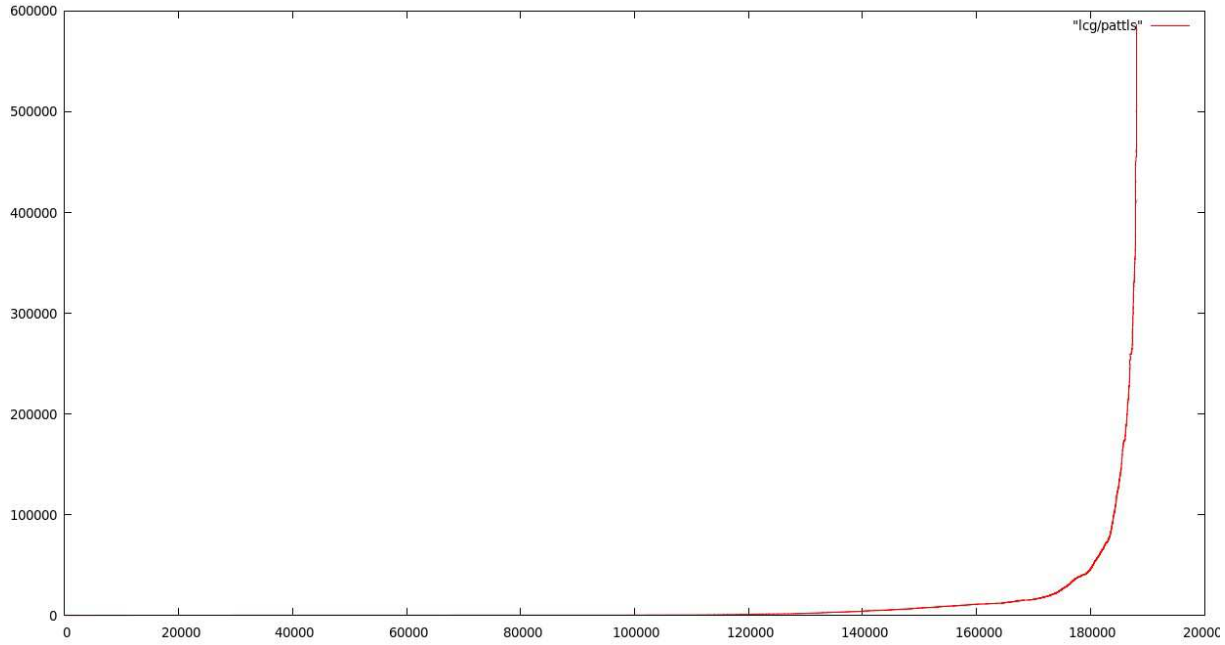

Figure 5: Job running times on the LCG platform in seconds, sorted.

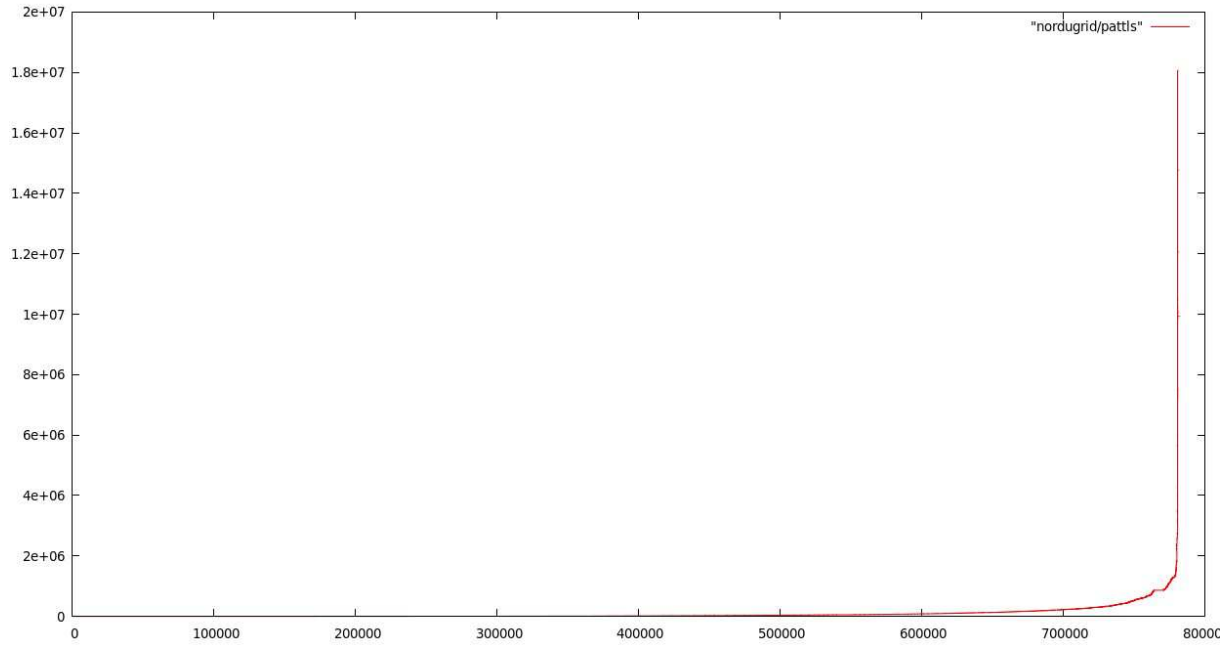

Figure 6: Job running times on the NorduGrid platform in seconds, sorted.

4 Experimental Results

In all our experiments we have used a time unit of 100 seconds and we have discretized the grid traces by this time unit. The plots of the grid traces and the predicted traces represent the total number of CPUs used by different jobs running in parallel in the time unit of 100 seconds. We have focused only on CPU usage as the information of memory usage was not available. Nevertheless, should the information of memory usage be available our approach can also be applied for its prediction.

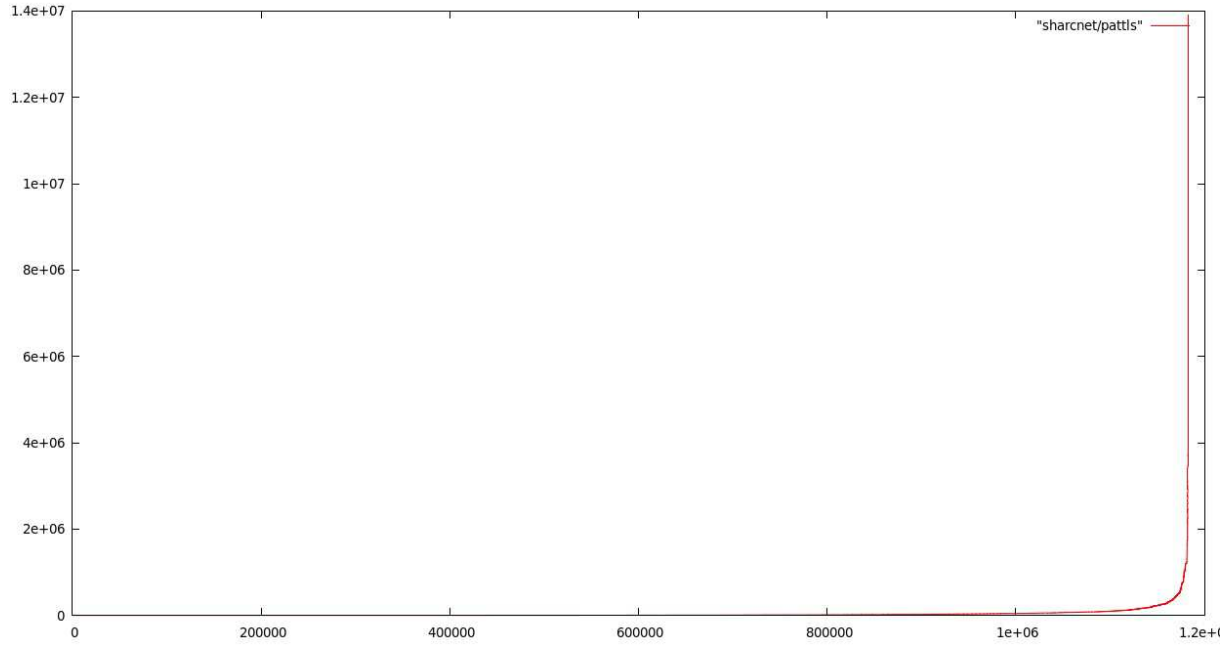

Figure 7: Job running times on the SHARCNET platform in seconds, sorted.

4.1 Data Sources

We have tested our auto-scaling approach with traces from three different research grids, each having its own usage particularities, with main differences in the frequency and amplitudes of changes in their overall usages.

LCG - Large Hadron Collider Computing Grid [1]

Here we find traces from several nodes from the computing grid associated to the Large Hadron Collider. Its behavior is mildly oscillatory and a plot of the total number of CPUs used in time slices of 100 seconds, discretized across time intervals of 100 seconds can be found in Figure 8.

NorduGrid [2]

Here we find higher amplitudes for oscillations as the grid is more heterogeneous than the previous. A plot of the total number of CPUs used in time slices of 100 seconds, discretized across time intervals of 100 seconds can be found in Figure 9.

SHARCNET [3]

SHARCNET has been described as a "cluster of clusters". Its volatility is very high and its amplitudes can reach surprising peaks. A plot of the total number of CPUs used in time slices of 100 seconds, discretized across time intervals of 100 seconds can be found in Figure 10.

4.2 Experiment setup

All the experiments use the server traces of the same form of input data as described above with time units of 100 seconds, and resource usage value consisting of the total number of CPUs used across the 100 seconds. A pattern length of 100 time units has been used for all the experiments (this is 100×100 seconds - approximately 2.7 hours of server time) and predictions are made for one time unit, this is 100 seconds, which is a little over 1 minute 30 seconds.

The results are displayed under the form of a set of standard metrics that include minimum, maximum, median, and average percentage and value difference between the prediction and the actual value.

[1] `http://lcg.web.cern.ch/LCG/`
[2] `http://www.nordugrid.org`
[3] `http://www.sharcnet.ca`

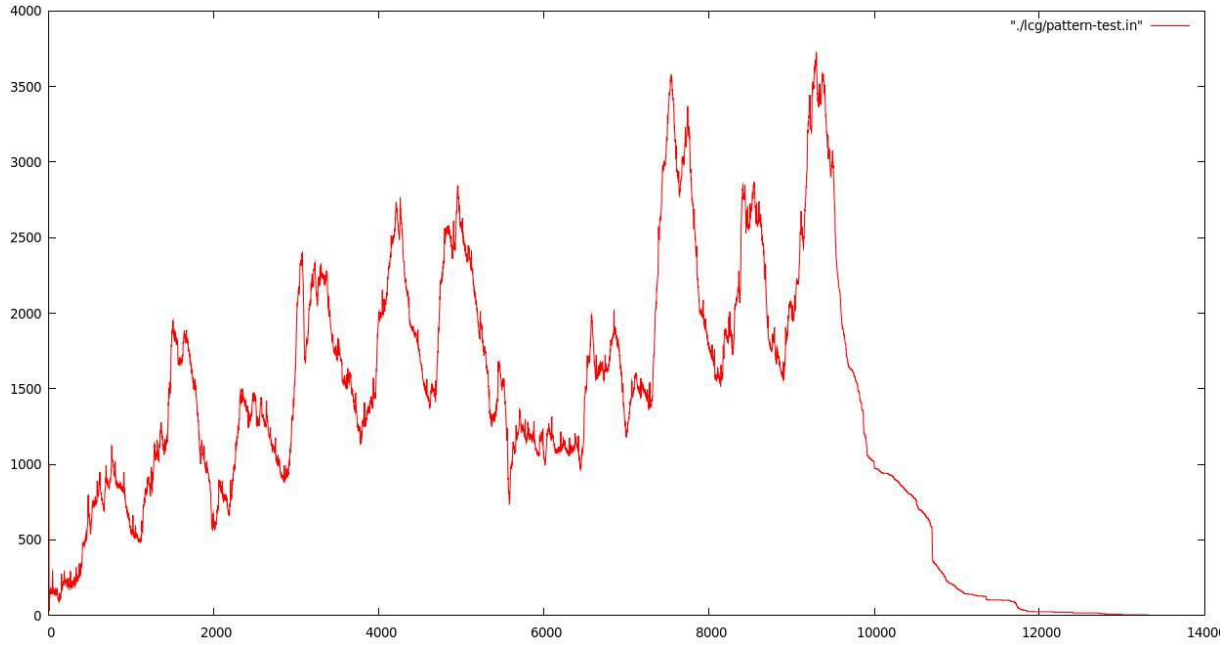

Figure 8: LCG - plot of the total number of CPUs used per time slice of 100 seconds versus time, discretized in slices of 100 seconds.

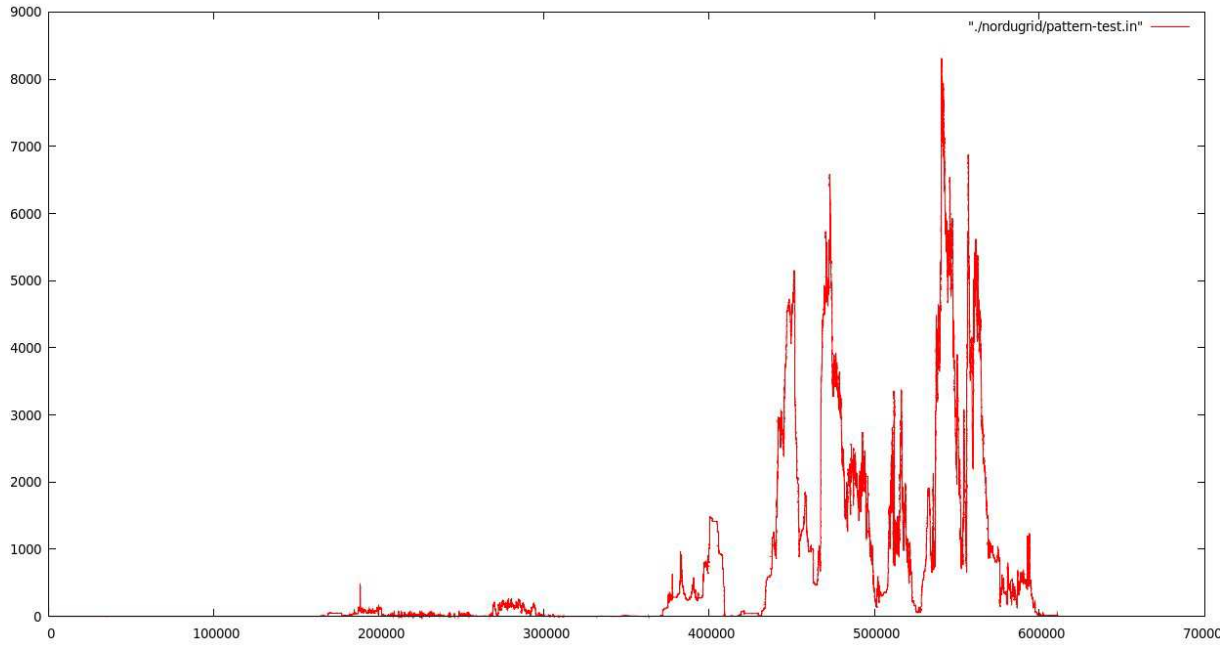

Figure 9: NorduGrid - plot of the total number of CPUs used per time slice of 100 seconds versus time, discretized in slices of 100 seconds.

A second set of metrics has also been used that allows the comparison to other existing auto-scaling algorithms. This metric was proposed and used by UCSB to compare the performance of three existing auto-scaling algorithms [11]: auto-regression of order 1, linear regression and the Rightscale democratic voting algorithm.

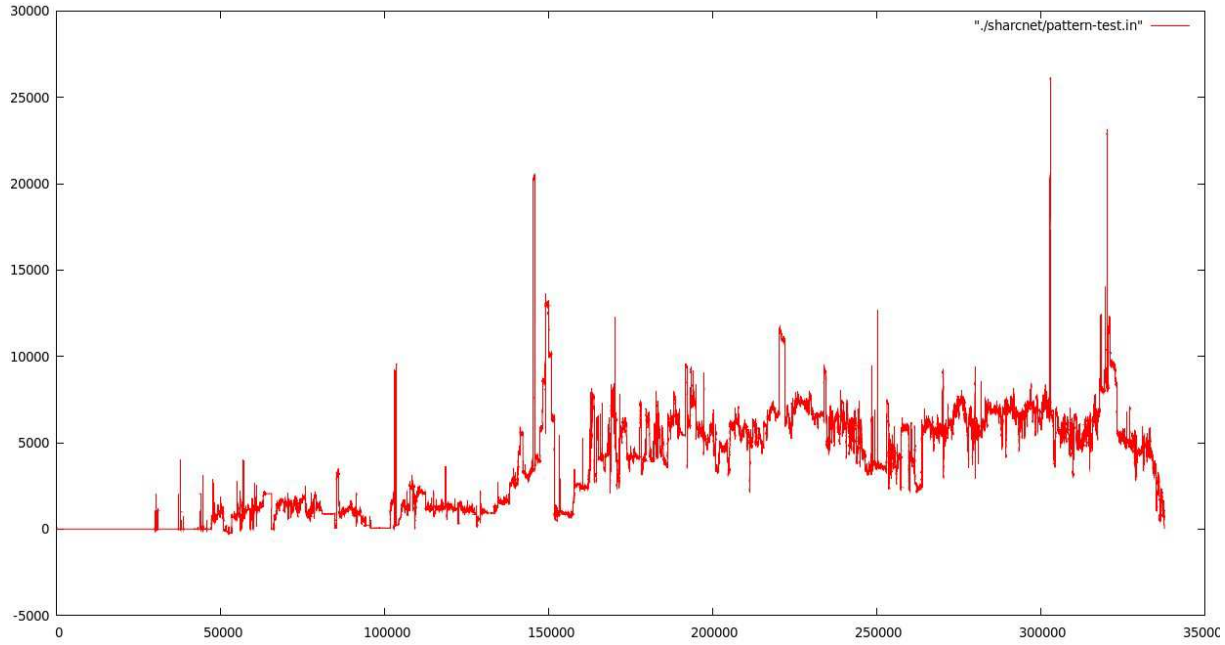

Figure 10: SHARCNET - plot of the total number of CPUs used per time slice of 100 seconds versus time, discretized in slices of 100 seconds.

We have also measured the average running time necessary for calculating one prediction. This has an impact on the practical usefulness of the prediction since it needs to be subtracted from the prediction time - which is 100 seconds - to calculate the effective prediction time.

We have used two versions of the metric proposed by the UCSB team:

- An instant score where we considered resource cost as being charged per fraction of an hour, although this is not the case in current cloud providers
- A second score where we take the maximum prediction over the course of an hour and use that as static provisioning for the whole hour

4.3 Results

Predicting LCG with LCG as historic data We have done a self-prediction test by using LCG as historic data with the purpose of predicting LCG itself. When filtering out potential pattern candidates, exact matches have been ignored, since the pattern itself is a piece of the historic data. The results of this experiment can be found in Table 4. Figure 11 shows a zoom-in of the actual value of resource usage in the LCG platform and the predicted resource usage.

Predicting NorduGrid with LCG as historic data

We have experimented with using traces from a different grid which is close to the one we are trying to predict. In the current test case, we have tried to predict NorduGrid workloads by using LCG as historic data. The experiment's results can be seen in Table 5. A zoom into the plots of the actual resource usage and the predicted usage is shown in Figure 12.

Predicting LCG with NorduGrid as historic data

We have experimented with the symmetric of the previous experiment in trying to predict LCG workloads by using NorduGrid as historic data. The results are shown in Table 6. A zoom into the plot of the actual resource usage of the platform and the predicted resource usage is shown in Figure 13.

Predicting SHARCNET with NorduGrid as historic data We have also experimented the behavior of the algorithm when using historic data that does not have a high similarity to the workload that is being

Metric	Value
Minimum percentage difference (%)	0.0
Minimum value difference	0.0
Maximum percentage difference (%)	53.4
Maximum value difference	220.97
Median percentage difference (%)	1.0
Median value difference	8.9
Average percentage difference (%)	1.749
Average value difference	15.33
UCSB metric (maximum per 1 hour)	5.44
UCSB metric (instantaneous)	-11.08
Average runtime for one instance (milliseconds)	41.734

Table 4: Results of predicting LCG with LCG as historic data. The top values represent the actual and the percentage difference between actual and predicted CPU usage for the platform, across time slices of 100 seconds. The middle two values represent the score of the prediction, according to the metric proposed by UCSB.

Metric	Value
Minimum percentage difference (%)	0.0
Minimum value difference	0.0
Maximum percentage difference (%)	1146.00
Maximum value difference	435.2
Median percentage difference (%)	1.74
Median value difference	0.26
Average percentage difference (%)	35.38
Average value difference	5.739
UCSB metric (maximum per 1 hour)	27.68
UCSB metric (instantaneous)	23.06
Average runtime for one instance (milliseconds)	162.949

Table 5: Results of predicting NorduGrid with LCG as historic data The top values represent the actual and the percentage difference between actual and predicted CPU usage for the platform, across time slices of 100 seconds. The middle two values represent the score of the prediction, according to the metric proposed by UCSB.

Metric	Value
Minimum percentage difference (%)	0.0
Minimum value difference	0.0
Maximum percentage difference (%)	100.0
Maximum value difference	217.12
Median percentage difference (%)	1.2
Median value difference	13.98
Average percentage difference (%)	7.32
Average value difference	18.46
UCSB metric (maximum per 1 hour)	3.43
UCSB metric (instantaneous)	-10.71
Average runtime for one instance (milliseconds)	514.956

Table 6: Results of predicting LCG with NorduGrid as historic data. The top values represent the actual and the percentage difference between actual and predicted CPU usage for the platform, across time slices of 100 seconds. The middle two values represent the score of the prediction, according to the metric proposed by UCSB.

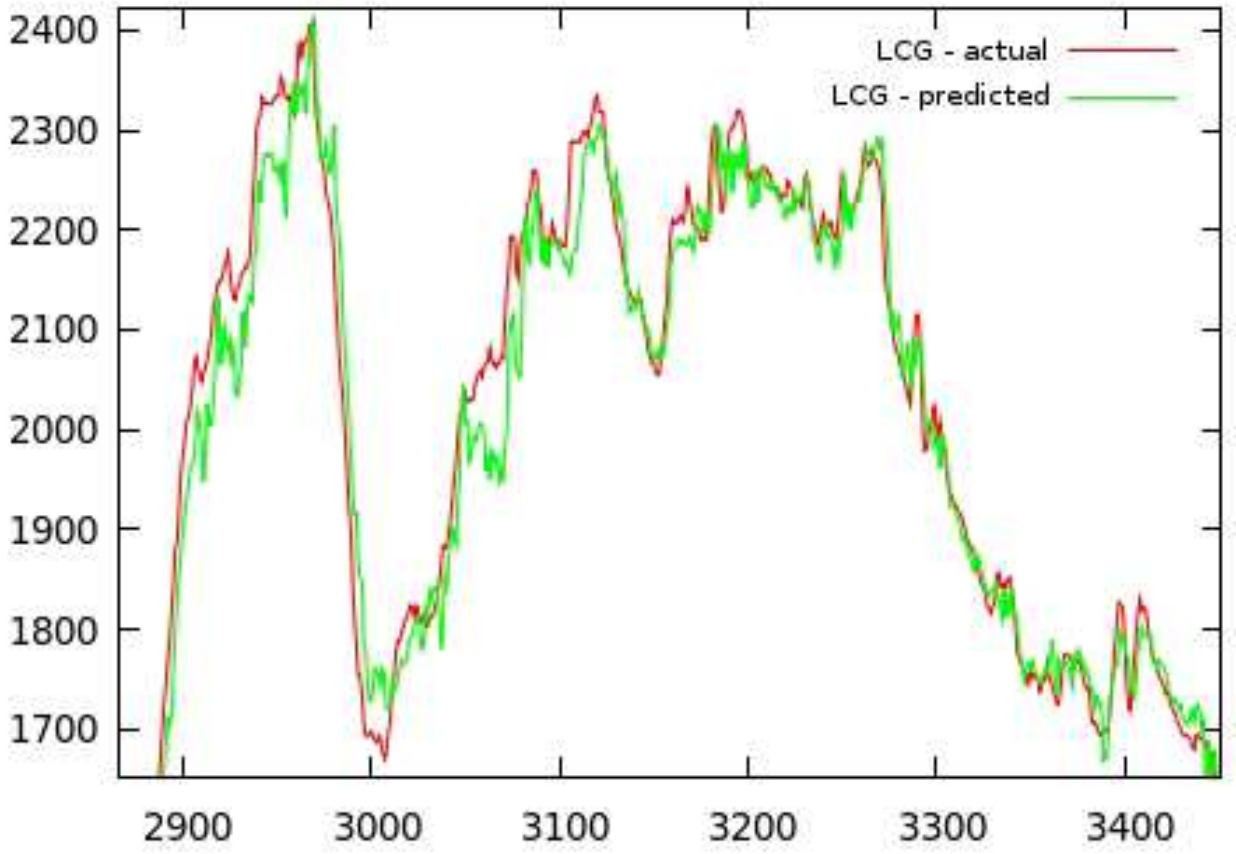

Figure 11: Zoom into the plot of CPUs used in time slices of 100 seconds versus time in units of 100 seconds for the LCG platform's actual resource usage (shown in red) and predicted resource usage (shown in green).

Metric	**Value**
Minimum percentage difference (%)	0.0
Minimum value difference	0.0
Maximum percentage difference (%)	528.03
Maximum value difference	5.64E17
Median percentage difference (%)	0.9
Median value difference	11.26
Average percentage difference (%)	375.65
Average value difference	4.03
UCSB metric (maximum per 1 hour)	-3.23
UCSB metric (instantaneous)	-2.06
Average runtime for one instance (milliseconds)	528.418

Table 7: Results of predicting SHARCNET with NorduGrid as historic data The top values represent the actual and the percentage difference between actual and predicted CPU usage for the platform, across time slices of 100 seconds. The middle two values represent the score of the prediction, according to the metric proposed by UCSB.

predicted. In our experiment, we have used NorduGrid traces as historic data when trying to predict SHARCNET traces. The results of this experiment are available in Table 7.

An analysis of the results reveals that this is a feasible approach to auto-scaling. It is clear that the algorithm yields better results when the set of historic data that is used has a similarity to the signal that is being predicted. This similarity is influenced by several parameters that constitute the domain of the server

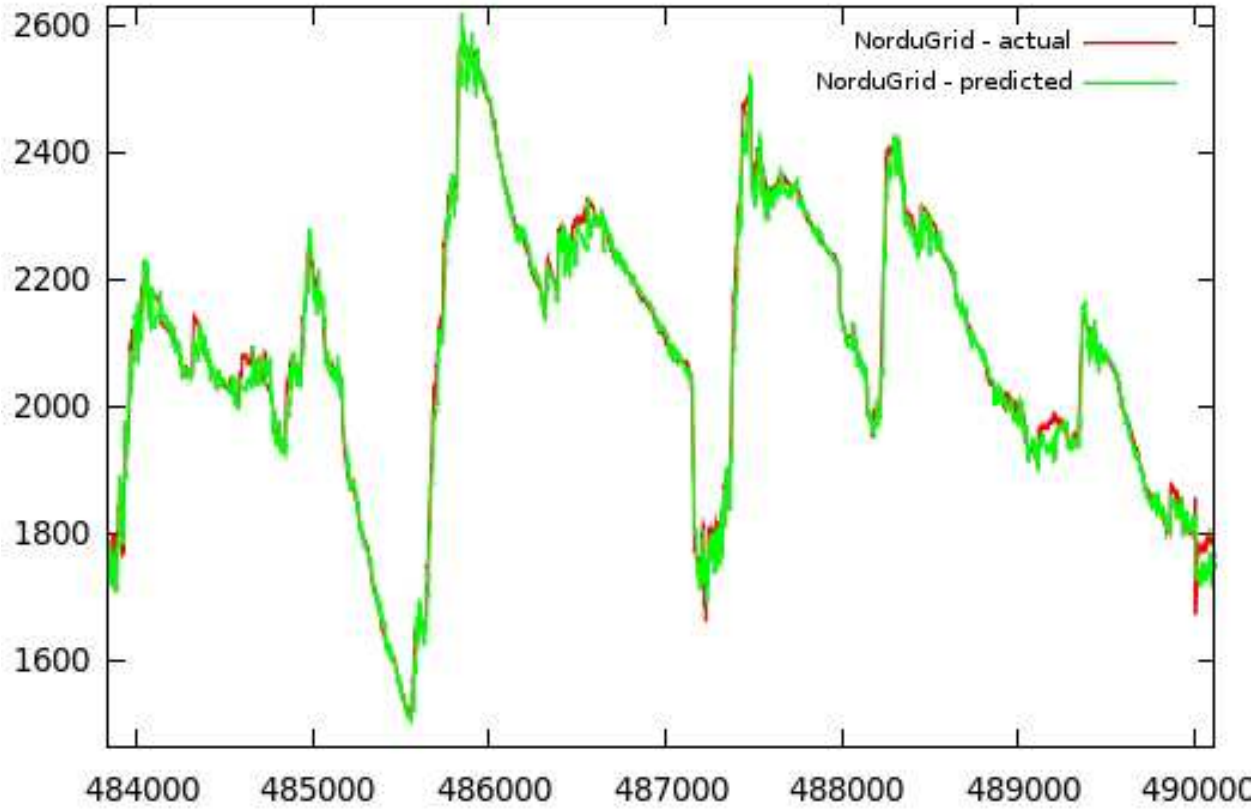

Figure 12: Zoom into the plot of CPUs used in time slices of 100 seconds versus time in units of 100 seconds for the NorduGrid platform's actual resource usage (shown in red) and predicted resource usage (shown in green) by using LCG as historic data

Pattern / data length	100.0%	50.0%	25.0%	12.5%
1000	-18.99	-36.37	-57.83	-97.37
500	-9.43	-19.97	-23.47	-43.06
100	5.44	3.32	4.05	4.05
50	9.41	9.6	8.48	8.21
25	10.67	11.11	12.62	11.79

Table 8: Score given by the UCSB metric (maximum per one hour) for predicting LCG with LCG as historic data and by varying the length of the pattern used for prediction and the length of the set of historic data. whose load is being predicted. It follows from the obtained results that data from the same domain can easily be used to predict one-another.

The time necessary for computing one prediction instance has proven in practice to be low relative to the prediction time.

Predicting LCG with LCG as historic data and varying pattern lengths and historic data lengths Although we cannot show that the algorithm yields the best results, we can show that its results improve as we increase the size of the historic data and as we find the best pattern length to take into consideration when predicting. The tables below illustrate results when varying the pattern length and the length of the historic data used for prediction. We have varied the historic data from 100% - the full set, to 50%, 25% and 12.5% of the set. The pattern length has also been varied from 1000 time units to 500, 100, 50 and 25.

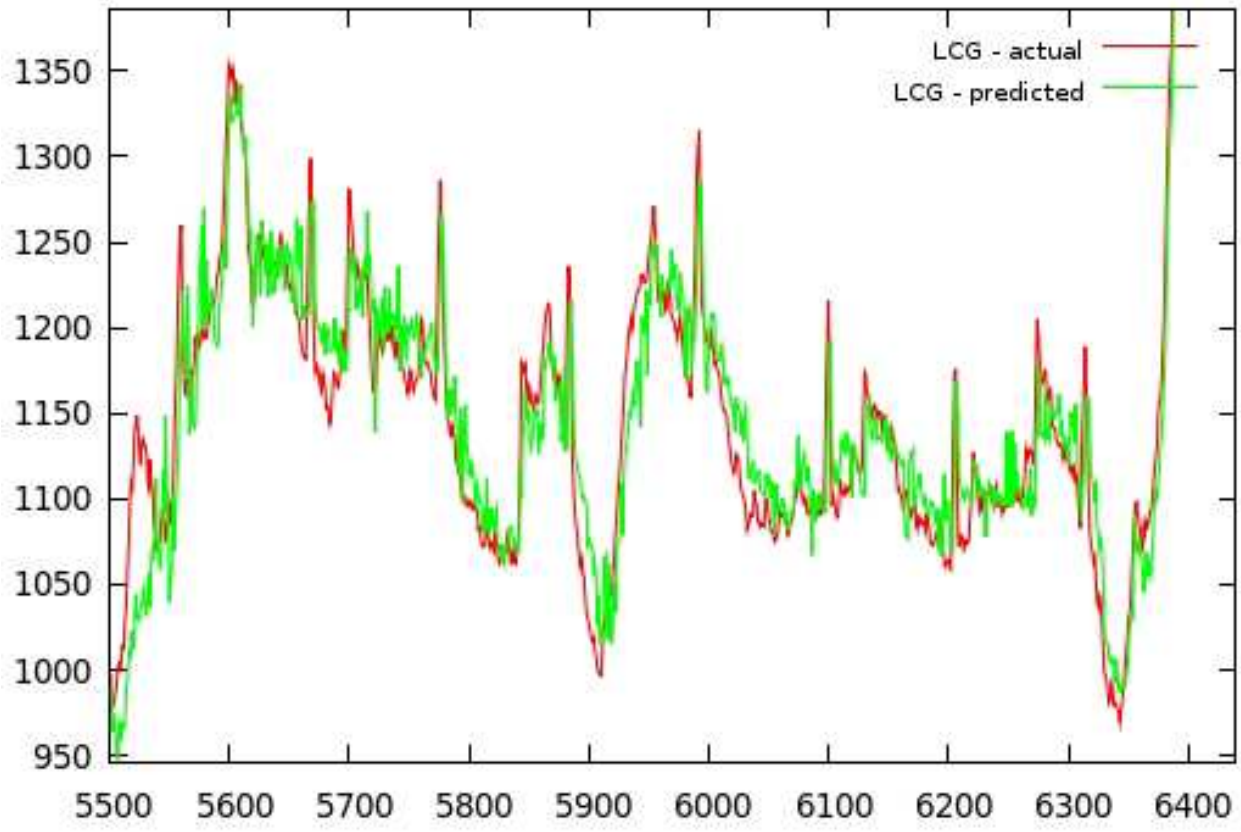

Figure 13: Zoom into the plot of CPUs used in time slices of 100 seconds versus time in units of 100 seconds for the LCG platform's actual resource usage (shown in red) and predicted resource usage (shown in green) by using NorduGrid as historic data.

Pattern / data length	100.0%	50.0%	25.0%	12.5%
1000	-38.96	-59.57	-79.25	-103.57
500	-31.36	-38.54	-45.88	-63.18
100	-11.08	-13.81	-16.49	-15.44
50	-4.54	-4.83	-7.22	-8.23
25	-0.14	-0.3	-0.05	-1.36

Table 9: Score given by the UCSB metric (instant) for predicting LCG with LCG as historic data and by varying the length of the pattern used for prediction and the length of the set of historic data.

Table 9 contains the results of the experiment when calculating the metric proposed in [11] and using instant values for the the number of virtual resources. Table 8 contains results of applying the previous metric by using the maximum across each hour as reference point for virtual resources and cost. There is a clear tendency for the prediction score to improve as the set of historic data increases in size. There is another tendency for the prediction score to increase as the length of the predicted pattern decreases. This is obvious if we take into consideration that a smaller pattern length corresponds to more identified pattern candidate, yet we do not recommend the usage of considerably small pattern lengths as this makes the result be more and more independent of the server usage trend.

The reader will note that in our experiments we have considered only CPU usage as measure and prediction target. In a Cloud environment, a virtual resource usually has more characteristics associated to it than just CPU power. In particular, memory usage is one of the most notable characteristics. Our approach can

also be used to have a prediction of the memory usage if the server traces also contain information about past memory usage. With predictions for both memory and CPU usages, the scaling component of the Cloud Client should be able to more accurately decide the characteristics of the virtual resources that are to be instantiated or released. The topic of making a good scaling decision both in direction and in virtual machine characteristics is an interesting topic of research, yet it is beyond the scope of the current work.

5 Conclusions and future work

One of the most important benefits of Cloud Computing is the ability for a Cloud Client to adapt the number of resources used based on its actual use. This has great implications on cost saving as resources are not paid for when they are not used. Dynamic scalability is achieved through virtualization. The downside of virtualization is that it has a non-zero setup time that has to be taken into consideration for an efficient use of the platform. It follows that an accurate prediction method would greatly aid a Cloud Client in making its auto-scaling decisions.

In this paper, a new resource usage prediction algorithm is presented. It uses a set of historic data to identify similar usage patterns to a current window of records that occurred in the past. The algorithm then predicts the system usage by interpolating what follows after the identified patterns from the historical data. Experiments have shown that the algorithm has good results when presented with relevant input data and, more importantly, that its results can improve by increasing the historic data size. This makes the evaluation of the algorithm be context dependent.

As future work directions we will be looking into ways that a relevant set of historic data can be composed for a particular application domain.

References

[1] M. Parashar N. Gnanasambandam A. Quiroz, H. Kim and N. Sharma. Towards autonomic workload provisioning for enterprise grids and clouds. In *Proceedings of the 10 th IEEE/ACM International Conference on Grid Computing (Grid 2009)*, pages 50 – 57, 2009.

[2] William I. Chang and Thomas G. Marr. Approximate String Matching and Local Similarity. In *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 259–273, London, UK, 1994. Springer-Verlag.

[3] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms, Chapter 32: String Matching*. McGraw-Hill Higher Education, 2001.

[4] Mark Crovella and Azer Bestavros. Explaining world wide web traffic self-similarity. Technical report, Boston, MA, USA, 1995.

[5] Nikolaos Doulamis, Anastasios Doulamis, Antonios Litke, Athanasios Panagakis, Theodora Varvarigou, and Emmanuel Varvarigos. Adjusted fair scheduling and non-linear workload prediction for qos guarantees in grid computing. *Computer Communications*, 30(3):499 – 515, 2007. Special Issue: Emerging Middleware for Next Generation Networks.

[6] Fermín Galán, Americo Sampaio, Luis Rodero-Merino, Irit Loy, Victor Gil, and Luis M. Vaquero. Service specification in cloud environments based on extensions to open standards. In *COMSWARE '09: Proceedings of the Fourth International ICST Conference on COMmunication System softWAre and middlewaRE*, pages 1–12, New York, NY, USA, 2009. ACM.

[7] GrenchMark. http://grenchmark.st.ewi.tudelft.nl.

[8] Large Hadron Collider Computing Grid. http://lcg.web.cern.ch/lcg/.

[9] Alexandru Iosup, , Ru Iosup, Dick H. J. Epema, Jason Maassen, and Rob Van Nieuwpoort. Synthetic grid workloads with ibis, koala, and grenchmark. In *In Proceedigs of the CoreGRID Integrated Research in Grid Computing*, 2005.

[10] Alexandru Iosup, Dick H. J. Epema, Carsten Franke, Alexander Papaspyrou, Lars Schley, Baiyi Song, and Ramin Yahyapour. On grid performance evaluation using synthetic workloads. In *JSSPP'06: Proceedings of the 12th international conference on Job scheduling strategies for parallel processing*, pages 232–255, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] Patricio Jara Jeff Browne Jonathan Kupferman, Jeff Silverman. Scaling Into The Cloud. 2009.

[12] Hashim Mohamed and Dick Epema. Koala: a co-allocating grid scheduler. *Concurr. Comput. : Pract. Exper.*, 20(16):1851–1876, 2008.

[13] NorduGrid. http://www.nordugrid.org.

[14] Radu Prodan and Vlad Nae. Prediction-based real-time resource provisioning for massively multiplayer online games. *Future Generation Computer Systems*, 25(7):785 – 793, 2009.

[15] SHARCNET. http://www.sharcnet.ca.

[16] TUDelft University The Grid Workloads Archive. http://gwa.ewi.tudelft.nl/pmwiki/pmwiki.php.

[17] Rob V. van Nieuwpoort, Jason Maassen, Gosia Wrzesińska, Rutger F. H. Hofman, Ceriel J. H. Jacobs, Thilo Kielmann, and Henri E. Bal. Ibis: a flexible and efficient java-based grid programming environment: Research articles. *Concurr. Comput. : Pract. Exper.*, 17(7-8):1079–1107, 2005.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399