A MODEL FOR CERTIFYING AND COSTING OF THE RESOURCE CAPABILITIES OF UTILITY COMPUTING RESOURCE CENTERS

Saif Faruqui, Ernesto Gomez, Yasha Karant, Keith Schubert
Department of Computer Science
And
Institute for Applied Supercomputing
California State University San Bernardino
5500 University Parkway
San Bernardino, CA 92407-2397

Contact author: Saif Faruqui sfaruqui@csci.csusb.edu (909)496-3582

This work was supported by the US National Science Foundation under award CISE 9810708, patent application in process.

ABSTRACT: *One of the many issues in the nascent field of utility computing is identification, on-the-fly, of available resources at different participating resource centers. Another fundamental issue is that of quantifying the available assets and maturity of a resource center organization with a view to compare different centers and select the centers best matching a user organizations requirements. This paper addresses the issues of assessment, certification, and costing of resource capabilities at utility computing resource centers. The various technical and business elements of a utility computing resource center are identified. With each of these elements a list of related factors is identified, that contribute to the cost of the element or can be used to assess and certify the capability of that resource. The certification factors are published in a certificate that a user can use to identify a center. The costing factors identified are placed in a matrix, and mathematically manipulated to arrive at a block diagonal matrix. This matrix can then be used, using a method developed by the authors (patent application in process), to arrive at a costing model. This model is flexible enough to accommodate different configurations of resource requirements by a user organization. Based on the set of resources required, the duration, and configuration a price or pricing model can be arrived at for that user. The detailed method in the process of being patented and thus is not disclosed here, for privacy purposes.*
Keywords: *Utility Computing, Pricing, Model, Factors, Certification.*

1. INTRODUCTION

One of the most exciting and promising areas of computer science presently is the utility computing paradigm. This envisages the deployment and use of computing resources in a utility model. In this model a number of computing resource centers will be hooked up to a compute grid. Compute power consumers will be able to use variable amounts of compute resources; storage, memory, compute cycles, bandwidth and special instruments, based on their requirements at the time, and pay only for what they use. A central issue in this model is the certification of resource centers, and further, specific subsets of resources in each center. This will enable resource consumers, and task allocation agents to take decisions on the use of a particular resource center. It will also help resource center owners to decide pricing structure for their services.

This paper proposes a model of certification, to certify utility computing resource centers -- outsourced and in-house utility computing resources available on demand. Pricing for resources used by a specific job will be determined by "market" forces. An important input is the prediction for resources required by a set of tasks submitted for simultaneous execution to a resource center, based on performance requirements. This will require the separate development of metrics for the same.

2. PHYSICAL ASSETS

Quantify capacity and performance of a particular resource center, for all the following resources. Quantify capacity available and performance predictions at a particular point in time, resource wise. This dimension is more relevant for on-the-fly decision making, for resource center users to decide for or against submitting a task to it.

2.1 FLOPS (floating point processing power)

The processing power available at a specific resource center is measured in FLOPS, or floating point operations per second. The reason for choosing this unit is that it is widely used to measure processing power of supercomputers, which are high-performance computers, like grid based distributed computing systems. But this unit has a drawback in that it is specific to programs on a specific processor. To compensate for this, the performance of each cluster with similar processors in a resource center will be measured for a set of standard applications. We propose to use a representative set of applications to obtain FLOPS data, including data mining, computational, visualization, and others that will be typical. These together represent the majority of categories of applications that will run on the computing resources in the resource center. These applications are the most widely used in the individual categories they represent.

2.2 TPS (database transactions)

For a job that needs to access a database to complete execution, the job assigner will need a measure of the maximum number of database transactions possible for each available kind of database. For this we propose to include a list of available databases and the maximum number of transactions possible for each database in our metric. Once a decision about which resource center to use has been initially taken, a client will store his data in the databases in that center. Subsequently, when a particular job tries to execute, it will check for the number of transactions, for the database of interest, that are presently available. This information will change dynamically as jobs are added and removed from the current list of the resource center.

2.3 Data communications throughput

This measure of the available resources of a resource center depends on network topology and congestion in the network at a resource center, and between the resource center and the job assigner. It is a dynamically changing quantity. Bandwidth requirements of all the presently executing jobs at a resource center will be predicted to calculate availability of bandwidth for an incoming resource. These predicted values will be modeled over time to check for congestion in the network on the resource center to decide if the incoming job can execute giving required performance. History of bandwidth availability will be maintained to predict bandwidth available between the allocation location and the resource center.

2.4 Mass storage capacity and mass storage rate to TPS engine, locally attached or over WAN

Total available secondary storage at a resource center will be a measure used to allocate jobs. The storage available could also include storage accessible over a WAN to the resource center. In the latter case available bandwidth would play a role in deciding which center to allocate a job to. Metrics published for this factor will be the total available storage, the storage per transaction, and the storage per kilobit/sec.

2.5 Specific architecture (processor, etc.)

Some jobs may require the availability of a specific hardware architecture to execute. This information will form another one of the parameters to decide on a resource center. All available hardware configurations will be published.

3. HUMAN ASSETS

This factor is useful for overall decision making, for an organization to decide whether to include a resource center in it's set of centers. Important attributes of the human assets are maturity, experience and quantifiable capability. A critical skill required in these human assets is the ability to forecast resource requirements for multiple tasks, running simultaneously, for a specific performance requirement. An important measure of the human resource of a utility computing organization will be the SEI P-CMM certification [3]. If the organization does have the certification then the level of P-CMM it has achieved will be the metric used for human resource measurement.

3.1 Management

Experience of management in running a mission critical center will be measured. This will serve as a measure of the confidence in management a potential client will have. A formula to assign an over all rating to the management of a resource center, on a scale will be used. This will take into consideration past experience in running mission critical computing, data or control centers.

3.2 Systems administration

The most important human element of a resource center are the systems administrators. A system for assessing the capabilities of these professionals will assign an overall rating to this factor. It will take into account certifications, experience, past record and education. A standardized questionnaire will be given to each of the system administrators in the organization to assess them followed by group interviews and personal interviews.

3.3 Development programmers

The development programmers will also be assessed on the same factors using the same method as the system administrators. The only difference will be the questionnaire they will be given.

4. SOFTWARE AND LICENSES

This dimension measures the software capability of the resource center. It also addresses legal compliance issues, both national and international. It is a factor in both overall and on-the-fly decision making. This is because; license possession may be an issue for on-the-fly decisions. For example, if a given user of the resource center needs to use more instances of a software package than he has done before, he will need information on licensing before taking a decision.

4.1 OS vender (HP-UX, AIX, MS Windows Server, etc.)

The OS used should allow multitasking and parallel execution. At any given time multiple applications could be running on one instance of the OS. Since the user will decide on a given center based on performance guarantees, the OS should assure a minimum level of performance on different hardware configurations.

4.2 ISV vendor (e.g., Oracle, SAP, PeopleSoft, Adobe, Quark, etc.)

The licensing mechanism used by software vendors will need to take into consideration varying levels of usage of their software package over a period of time. The license agreement should allow as many instances as needed by the user and charge the resource center accordingly.

5. SECURITY

Security is a central issue in the utility computing model. Tasks will be submitted to resource centers after matching the security requirement of each task with the security rating of that center. This can be further refined to allocate a security rating to sub-parts of a resource center, thus further differentiating individual sets of resources at a site. Security will also be an important factor in billing rates. The security rating of a resource center will be decided by grading it on a predefined set of parameters. The following parameters will be used for our initial set:

1. Completion of a thorough Security Policy
2. Implementation of a complete Incident Management procedure
3. Completion of a Risk Assessment report
4. Completion of a Threat / Vulnerability Analysis
5. Development of an audited Security Architecture
6. Appropriate deployment of Network Intrusion Detection systems
7. Anti Viral Software Policy & Implementation
8. Network Architecture and Configuration policy
9. Establishment & Conduction of rigorous Auditing procedures
10. Staff screening
11. Authentication mechanisms
12. Authorization mechanisms
13. Repudiation mechanisms

5.1 Indemnification

Each task submitted to a center will need to have a monetary value attached to it. This will enable tasks to be insured against loss or theft. The specific value of a task or of data can be decided using standard tables and negotiations held in advance.

6. OVERALL AUTHORITY TO CERTIFY 1, 2, 3, 4, AND 5

The entity certifying a given resource center should have the authority to certify a resource center. This could include the authority to grant an ISO, SEI CMM (P-CMM), Six Sigma, etc., certification.

7. CONCLUSION: THE METHOD

A method for costing of services at the utility computing resource center has been developed. This method will help the center owner to decide on a configuration for his center and the pricing structure. This will also help center clients to decide on which center to buy resources from. Additionally, the certification parameters mentioned in each of the factors above will be assigned values and published. This certificate can also be used to assign a level to the resource centers using a method similar to the SEI-CMM model. And, these will be the guiding factors for certifying authorities to certify utility computing resource centers.

The costing factors can be divided into groups of factors that are closely related to each other, with each factor assigned a code. The following groups have been identified:

1. Database transactions: a) Cost/Transaction/Database (c00); b) Database administration (c01); c) Database software (c02).
2. Data communication: a) Bandwidth (c10); b) Redundancy (c11); c) Network administration (12); d) Network software (c13); e) Security software (c14).
3. Mass storage: a) Security software (c20); b) Storage (c21); c) Storage software (c22); d) Human resources (c23); e) Redundancy (c24).
4. Common costs: a) Common human assets (c30); b) Common management and overhead (c31); c) Common security (c32); d) Common indemnification (c33).
5. Hardware architecture: a) Unit cost of each configuration of similar processors (c40); b) Power requirements (c41).
6. OS and software: a) Licensing (c50); b) System administration (c51).
7. Indemnification: a) Indemnification cost per job (c60).

All these cost factors are considered at a fixed point in time. Each of them can be specified in the form cXY(f), which is a particular cost, denoted by X and Y, as a function of fixed time.

The resulting matrix is:

C: = $| c_{ij} |$ (i = 0,6, j=0,4), with the i and j representing the factors above.

Each cost factor in a related group (same X in cXY) will be assigned a relative weight, wji. The summation of the product of weight and cost in each group will give us Ç(j) = sigma (wji*cji) = WC. W being the eigenvalues of the Ç(j) matrix. The W matrices for each group of factors will then be placed in a block diagonal matrix, with the assumption of mutual independence between blocks. This matrix can be used by a utility center owner to arrive a cost estimate for a particular resource set requirement.

References

[1] Rothenberger, M. A., Dooley, K. J., Kulkarni, U. R., and Nada, N. "Strategies for Software Resue: A Principal Component Analysis of Reuse Practices", *IEEE Tans. Software Eng.*, 2003, 9: 825-837.

[2] Hoffman, T. Com*HP takes new pricing path for utility-based computing,* Online Article, ComputerWorld, http://www.computerworld.com/managementtopics/management/itspending/story/0,10801,81522,00.html.

[3] CMU SEI, http://www.sei.cmu.edu/sei-home.html, Online Documents on SEI CMM.