Secure Data Processing over Hybrid Clouds

Vaibhav Khadilkar [#1], Kerim Yasin Oktay [*1], Murat Kantarcioglu [#2] and Sharad Mehrotra [*2]

[#] *The University of Texas at Dallas*

[1] vvk072000@utdallas.edu, [2] muratk@utdallas.edu

[*] *University of California, Irvine*

[1] koktay@uci.edu, [2] sharad@ics.uci.edu

Abstract

A hybrid cloud is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability [1]. The emergence of the hybrid cloud paradigm has allowed end-users to seamlessly integrate their in-house computing resources with public cloud services and construct potent, secure and economical data processing solutions. An end-user may be required to consider a variety of factors, which include several hybrid cloud deployment models and numerous application design criteria, during the development of their hybrid cloud solutions. Although a multitude of applications could be developed through a combination of the aforementioned deployment models and design criteria, the common denominator among these applications is that they partition an application's workload over a hybrid cloud. Currently, there does not exist a framework that can model this workload partitioning problem such that all the previously mentioned factors are considered. Therefore, in this paper we present our vision for the formalization of the workload partitioning problem such that an end-user's requirements of performance, data security and monetary costs are satisfied. Furthermore, to demonstrate the flexibility of our formalization, we show how existing systems such as [2] [3] can be derived from our general workload partitioning framework through an instantiation of the appropriate criteria.

1 Introduction

The emergence of cloud computing has created a paradigm shift within the IT industry by providing users with access to high quality software services (SaaS), robust application development platforms (PaaS) and sophisticated computing infrastructures (IaaS). Furthermore, the utilization of a pay-as-you-use pricing model for usage of cloud services is a particularly inviting feature for users, since it allows them to significantly lower their initial investment cost towards acquiring a cloud infrastructure. A hybrid cloud is a particular cloud deployment model that is composed of two or more distinct cloud infrastructures (private, community, or public) that remain autonomous entities, but are interleaved through standardized or proprietary technology that enables data and application portability [1]. A growing number of organizations have turned to such a **hybrid cloud model** [4] [5], which allows them to seamlessly integrate their private cloud infrastructures with public cloud service providers. A hybrid cloud model enables users to process organization-critical tasks on their private infrastructure while allowing repetitive, computationally intensive tasks to

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

be outsourced to a public cloud. Moreover, adopting a hybrid cloud model increases throughput, reduces operational costs and provides a high level of data security.

There are several flavors of hybrid cloud deployment models that are available to a user. One of the choices is the **Outsourcing model**, in which users outsource all tasks to public cloud service providers. Additionally, a user's private cloud infrastructures are mainly used to perform post-processing operations such as filtering incorrect results or decrypting encrypted results. Some features of the Outsourcing model include ease of application development and deployment, reduction in financial expenditure and guaranteed levels of performance through the use of Service Level Agreements (SLA's). A second choice is the **Cloudbursting model**, in which users primarily use an in-house cloud infrastructure for deploying applications and use public cloud services to mitigate sudden bursts of activity associated with an application. A Cloudbursting model provides users with several advantages such as adaptability to changing computational capacity requirements and cost savings through efficient use of public cloud resources. A third choice is the **Hybrid model**, in which users could host applications that operate over sensitive information on a private infrastructure while outsourcing less critical applications to a public cloud service provider. Such a deployment model offers a higher level of throughput, an enhanced data security and an overall reduction in financial costs. Given the various choices we have just described, a user needs to make a mindful selection of a particular deployment model based on their application requirements.

In addition to the deployment models presented above, a user also needs to consider a variety of criteria for designing a hybrid cloud application. The single most important criterion is **Performance**, since any design solution must strictly adhere to a user's performance requirements. The performance of a hybrid cloud application depends on several criteria such as the data model used to capture information and the data representations used to store data on a public cloud. The second important criterion that merits a user's consideration is **Data Disclosure Risk**, since a hybrid cloud application outsources tasks, and implicitly data, to a public cloud, thereby creating a potential risk if the data is leaked. The data disclosure risk is dependent on factors such as the representation used to store data on a public cloud and whether a selected representation discloses information during data processing. The third important criterion that deserves a user's attention is **Resource Allocation Cost**, since an application's usage of cloud services leads to expenses that must be covered by an organizational budget. The resource allocation cost is contingent on the cloud vendor and type of services being commissioned. The last essential touchstone that a user should consider is **Private Cloud Load**, since for certain deployment models, namely Outsourcing and Cloudbursting, a user would necessarily want to limit the amount of processing that is performed on a private cloud. In practice, the load generated on a private cloud primarily depends on the model used to capture and process data. Given the multitude of design criteria we described, a user is required to make selections in a way that effectively addresses their performance, security and financial requirements.

In this paper, we begin by identifying the most notable criteria, which were briefly outlined earlier, that drive the design of an effective hybrid cloud solution. In addition, we also tabulate the applicability of these criteria to various cloud deployment models, which were introduced earlier. An observation to be made at this point is that, although a user is required to consider a variety of factors, such as several hybrid cloud deployment models as well as numerous application design criteria, the common denominator among any applications that are developed using the aforementioned factors is that they partition an application's workload over a hybrid cloud. In this paper, we formalize this workload partitioning problem as a mechanism for maximizing a workload's performance and we subsequently develop a framework for distributing an application's workload over a hybrid cloud such that an end-user's requirements with respect to performance, data disclosure risk, resource allocation cost and private cloud load are satisfied. We then describe how existing systems such as [2] [3] can be derived from the general workload partitioning framework through an instantiation of the appropriate parameters.

Our primary technical contributions are listed below:

- We identify the most significant criteria that drive the design of an effective hybrid cloud solution. In addition, we demonstrate the applicability of these criteria to various cloud deployment models.
- We formalize the workload partitioning problem as a mechanism for maximizing workload performance. Our formalization allows us to plug in various models for metrics that have the greatest impact towards the effectiveness of a hybrid cloud deployment model. In addition, our formalization allows an end-user to experiment with different levels of restrictions for public cloud usage, until they achieve the right mix of performance, security and financial costs.
- We demonstrate the flexibility of our formalization by showing how existing systems such as Sedic [2] as well as the work given in [3], henceforth referred to as Hybrid-I, can be derived from the general workload partitioning framework through an instantiation of the appropriate design criteria.

The rest of the paper is organized as follows: In Section 2 we present several key design criteria that we believe are essential towards the development of an effective hybrid cloud solution. Then, Section 3 presents a general formalization of the workload partitioning problem that is applicable to any hybrid cloud deployment model using the design criteria outlined in Section 2. After that, Section 4 describes how existing systems can be derived from the general workload partitioning framework based on a specification of concrete values for the appropriate design criteria. Finally, we describe our conclusions and future work in Section 5.

2 Design Criteria for Hybrid Cloud Models

In this section, we present a brief overview of the design criteria that provide the greatest contribution towards an effective hybrid cloud solution. Furthermore, Table 2 shows how these criteria are applicable to hybrid cloud models (*viz.* Outsourcing, Cloudbursting and Hybrid) as well as a Private-only cloud.

Performance: This criterion is the single most important one for the adoption of hybrid clouds, since a user would be willing to consider a cloud approach only if it meets their evolving performance requirements. In the context of hybrid clouds, there are several, mutually conflicting metrics that could be used to measure performance. These include, query response time and network throughput, among others. The performance of a hybrid cloud model is in turn dependent on several factors such as data model, sensitivity model, *etc*.

Data Disclosure Risk: This factor estimates the risk of disclosing sensitive data to a public cloud service provider, albeit in an appropriately encrypted form [6]. The risk is contingent on the the sensitivity and security models defined by a user. Furthermore, the risk could be measured using a simple metric such as the number of sensitive cells exposed to a public cloud [3] or a more complex analytical [7] or entropy-based [8] technique.

Resource Allocation Cost: This criterion measures the financial cost (in terms of $) engendered by the incorporation of some type of public cloud services into hybrid cloud models. The cost can be classified into the following two broad categories: (i) On-premise: This category measures the cost incurred in acquiring and maintaining a private cloud. (ii) Cloud: This category can be further sub-divided as follows: (a) Elastic: A user is charged only for the services they use (pay-as-you-use). (ii) Subscription: A user is charged a decided fee on a regular basis (fixed). The financial cost of an end-user's hybrid cloud model implementation is dependent on several factors such as the data model/query language, storage representation, *etc*.

Private Cloud Load: This touchstone estimates the load on a private cloud generated as a result of processing some part of a user's workload. This criterion is particularly appropriate in the context of the Outsourcing and Cloudbursting deployment models, where the goal of a user is to avoid processing any data or processing a small amount of data on a private cloud. The load on a private cloud could be measured using a variety of metrics such as workload response time, total number of I/O operations performed when a workload is processed, *etc*.

Observations: There are several observations to be made from the criteria we have listed above.

Table 2: Design Criteria and their Applicability to Cloud Models

Design Criteria	Private-only	Outsourcing	Cloudbursting	Hybrid
Performance	✓	✓	✓	✓
Data Disclosure Risk	×	✓	✓	✓
Resource Allocation Cost	On-premise	Cloud	Both	Both
Private Cloud Load	×	✓	✓	✓

- The different criteria are tightly coupled with one another, thus requiring a methodical selection process to successfully accomplish an end-user's requirements.
- The main distinguishing characteristic between the Outsourcing model *vs.* the Cloudbursting and Hybrid models is the Data Disclosure Risk. In the Outsourcing model, the disclosure risk is higher than the Cloudbursting and Hybrid models, since the entire dataset and workload are outsourced to a public cloud[1]. On the other hand, the disclosure risk in the Cloudbursting and Hybrid models can be configured as an adjustable parameter, thus causing the overall risk in these models to be lower than the Outsourcing model.
- Although the Cloudbursting and Hybrid models appear to overlap in terms of the criteria described above, there are two important differences between them: (i) In the Cloudbursting model, private cloud data is always replicated on a public cloud. The level of replication, *viz*, partial or full, is dependent on an end-user's choice. However, in the Hybrid model, an end-user decides whether data replication is performed. (ii) In the Cloudbursting model, computations are pushed to a public cloud only when the generated load begins to exhaust private cloud resources. In the Hybrid model, a user's preference dictates whether private cloud load is used as a criterion for distributing a workload.

3 Workload Partitioning Problem

In this section, we formalize the workload partitioning problem, WPP, for a hybrid cloud setting, using the design criteria we outlined in Section 2. The goal of WPP is to distribute a workload W, and implicitly a dataset R, over a hybrid cloud deployment model such that the overall performance of W is maximized. Additionally, the problem specification is bounded by the following constraints: (i) **Data Disclosure Risk**: The risk an end-user is willing to accept due to disclosure of sensitive data stored on a public cloud. (ii) **Public Cloud Resource Allocation Cost**: A user-defined upper bound on monetary costs, which limits the amount of public cloud services that could be leased for processing data. (iii) **Private Cloud Load**: The permissible capacity to which private cloud resources could be commissioned for processing data.

WPP **Definition**: Given a dataset R and workload W, WPP can be modeled as an optimization problem whose goal is to find a subset $W_{pub} \subseteq W$ of the workload, and implicitly a subset $R_{pub} \subseteq R$ of the dataset such that the overall performance of W is maximized.

$$
\begin{aligned}
&\text{maximize} && Performance(W, W_{pub}) \\
&\text{subject to} && (1)\ Risk(R_{pub}, Rep) \leq DISC\_CONST \\
& && (2)\ Pricing(R_{pub}, W_{pub}) \leq PRA\_CONST \\
& && (3)\ Load(W - W_{pub}) \leq LOAD\_CONST
\end{aligned}
$$

[1]Public cloud services such as Amazon S3 allow users to store data in an encrypted format at no additional monetary costs [9]. This facility ensures that data is protected when it is unused, however, the data is in cleartext form when it is brought into memory during processing, and hence is susceptible to memory attacks at this time [10].

where $DISC\_CONST$, PRA_CONST and $LOAD\_CONST$ denote the maximum admissible data disclosure risk, public cloud resource allocation cost and private cloud load as specified by an end-user. The general formalization of WPP given above extracts and presents the essential components of the workload partitioning problem in the context of various hybrid cloud deployment models. Furthermore, such a general framework allows us to construct several practical hybrid clouds by instantiating each of the criteria specified in Section 2 with different values. Additionally, a general specification enables us to systematically analyze the interdependence between the design criteria and thus assist users in making informed choices for the various criteria.

The general formalization of WPP includes a high-level mathematical definition of various metrics, namely performance, data disclosure risk, public cloud resource allocation cost and private cloud load, which collectively assist us in measuring the effectiveness of a hybrid cloud deployment model. This high-level definition needs to be further refined for a particular hybrid cloud variant based on the values specified for the various design criteria outlined earlier. Therefore, in the subsequent section, we present specific instantiations of the applicable metrics, as defined by Sedic and Hybrid-I.

4 Sample Variants of WPP for the Hybrid Cloud Deployment Model

In this section, we demonstrate the flexibility of our formalization by showing how existing systems such as Sedic [2] and Hybrid-I [3] can be derived from the general workload partitioning framework through a specification of concrete values for the appropriate design criteria we identified earlier.

Sedic: An inherent drawback to existing cloud computing frameworks, such as MapReduce, is their inability to automatically partition a computational task such that computations over sensitive data are performed on an organization's private cloud, while the remaining data is processed on a public cloud. The goal of Sedic is to address this drawback by enhancing the MapReduce framework with special features that allow it to partition and schedule a task over a hybrid cloud according to the security levels of the data used by the task.

The workload partitioning problem definition for Sedic [2] can be constructed by using the values given in Table 3 for the various design criteria. Note that, Sedic also uses the following specifications: (i) Data Model: Key-Value. (ii) Data Partitioning Model: None. (iii) Data Replication Model: Full replication of non-sensitive data to public cloud. (iv) Sensitivity Model: Sensitivity is defined at data-level using a labeling tool. (v) Security Model for Public Clouds: All sensitive data is sanitized to 0. (vi) Workload Model: A single MapReduce job.

Table 3: Design Criteria Specification for Sedic

Design Criteria	Specification
Performance	Overall Task Execution Time
Data Disclosure Risk	0, *viz.*, no sensitive data is exposed
Resource Allocation Cost	None
Private Cloud Load	Not considered

WPP **Definition for Sedic**: Since Sedic supports single MapReduce jobs, W can be modeled as a workload of tasks T, where a task is either a Map or Reduce task. Then, WPP can be defined as follows for Sedic: Given a dataset R and a task workload T, a variant of WPP for Sedic can be modeled as an optimization problem whose goal is to find subsets $T_{pub} \subseteq T$ and $R_{pub} \subseteq R$ such that the overall execution

time of T is minimized.

$$\begin{array}{ll} \text{minimize} & Performance(T, T_{pub}) \\ \text{subject to} & (1)\ Risk(R_{pub}, Rep) \leq DISC\_CONST \end{array}$$

where, as before, $DISC\_CONST$ denotes the maximum permissible data disclosure risk, which is 0 for Sedic, since no sensitive information can be leaked to a public cloud. In addition, the following observations can be made from the WPP definition of Sedic based on the specifications given above: (i) A data item $R_i \in R$ denotes either a Key or Value, since Sedic uses the Key-Value data model, no partitioning and a full data replication model. (ii) The set Rep consists of two representations, namely "plaintext" and "0", since Sedic sanitizes all sensitive data stored on a public cloud to 0.

We now provide specific instantiations of performance and data disclosure risk that suitably capture aspects of the metrics that are relevant to the problem domain modeled by Sedic.

Performance: As stated earlier, Sedic uses the *overall task execution time* of workload T, denoted as $ORunT(T, T_{pub})$, as an indicator of performance. Consequently, the objective function of WPP aims to minimize the overall execution time of a given task workload T. The execution time of tasks in T over a hybrid cloud, given that tasks in T_{pub} are executed on a public cloud can be represented as follows:

$$Performance(T, T_{pub}) = ORunT(T, T_{pub}) = \max \left\{ \begin{array}{l} \sum\limits_{t \in T_{pub}} runT_{pub}(t) \\ \sum\limits_{t \in T - T_{pub}} runT_{priv}(t) \end{array} \right.$$

Note that $T_{pub} \subseteq T$, otherwise it is undefined. Additionally, $runT_x(t)$ denotes the estimated running time of task $t \in T$ at site x where x is either a public ($x = pub$) or private ($x = priv$) cloud. In practice, a methodology such as that given in [11] can be used to estimate the running time of a task t as follows:

$$runT_x(t) = \left\{ \begin{array}{l} totalMapTime = \left\{ \begin{array}{ll} cReadPhaseTime + cMapPhaseTime + cWritePhaseTime, & \text{if } pNumReducers = 0 \\ cReadPhaseTime + cMapPhaseTime + cCollectPhaseTime + & \\ cSpillPhaseTime + cMergePhaseTime, & \text{if } pNumReducers > 0 \end{array} \right. \\ totalReduceTime = cShufflePhaseTime + cMergePhaseTime + cReducePhaseTime + cWritePhaseTime \end{array} \right.$$

where the semantics associated with the different variables used above are given in Table 4. An interested reader can refer to the technical report given in [11] for additional details.

Data Disclosure Risk: Sedic uses a data labeling tool to mark sensitive subsets, R_i, of a dataset R. Furthermore, Sedic sanitizes any marked out sensitive data, which needs to be stored on a public cloud, to 0. A combination of these two factors ensures that no sensitive data is exposed to a public cloud, *viz.* $Risk(R_{pub}, Rep) = 0$. Moreover, no sensitive information is leaked from a public cloud, since all sensitive values are sanitized to the same value, *viz.* 0. Finally, since no sensitive data is exposed to a public cloud, Sedic ensures that the data disclosure risk constraint, namely $DISC\_CONST$, which has a value of 0 for Sedic, is not violated.

Hybrid-I: A common characteristic across all hybrid cloud applications is that they partition the application's computational workload, and implicitly the data, over a hybrid cloud. However, a user has a multitude of computation partitioning choices based on their desired application requirements. Moreover, it is infeasible to construct applications over each of the possible computation partitioning choices. The goal of Hybrid-I is to formalize the computation partitioning problem over hybrid clouds such that an end-user's desired requirements are achieved. Additionally, Hybrid-I provides a dynamic programming solution to the computation partitioning problem, when the underlying workload consists of Hive queries and the dataset is assumed to be relational.

The workload partitioning problem definition for Hybrid-I can be constructed through an instantiation of the various design criteria using the values given in Table 5. Note that, Hybrid-I also uses the following

Table 4: Semantics of Variables used in the estimation of the running time of a task t

Variable	Semantics
$cReadPhaseTime$	The time to perform the Read phase in a Map task
$cMapPhaseTime$	The time to perform the Map phase in a Map task
$cCollectPhaseTime$	The time to perform the Collect phase in a Map task
$cSpillPhaseTime$	The time to perform the Spill phase in a Map task
$cMergePhaseTime$	The time to perform the Merge phase in a Map/Reduce task
$cShufflePhaseTime$	The time to perform the Shuffle phase in a Reduce task
$cReducePhaseTime$	The time to perform the Reduce phase in a Reduce task
$cWritePhaseTime$	The time to perform the Write phase in a Map/Reduce task
$totalMapTime$	The overall time to perform a Map task
$totalReduceTime$	The overall time to perform a Reduce task

specifications: (i) Data Model: Relational. (ii) Data Partitioning Model: Vertical. (iii) Data Replication Model: Partial replication of data to public cloud. (iv) Sensitivity Model: Attribute-level. (v) Security Model for Public Clouds: Bucketization [12]. (vi) Workload Model: Hive[2] queries in batch form.

Table 5: Design Criteria Specification for Hybrid-I

Design Criteria	Specification
Performance	Overall Query Execution Time
Data Disclosure Risk	No. of sensitive tuples exposed to Public cloud
Resource Allocation Cost	Cloud - Elastic
Private Cloud Load	Not considered

WPP **Definition for Hybrid-I**: Since Hybrid-I uses Hive queries in batch form, the workload W can be modeled as a set of Hive queries Q. Then, the WPP definition for Hybrid-I can be given as follows: Given a dataset R and a query workload Q, a variant of WPP for Hybrid-I can be modeled as an optimization problem whose goal is to find subsets $Q_{pub} \subseteq Q$ and $R_{pub} \subseteq R$ such that the overall execution time of Q is minimized.

$$\begin{aligned} \text{minimize} \quad & Performance(Q, Q_{pub}) \\ \text{subject to} \quad & (1)\ Risk(R_{pub}, Rep) \leq DISC\_CONST \\ & (2)\ Pricing(R_{pub}, Q_{pub}) \leq PRA\_CONST \end{aligned}$$

where, as before, $DISC\_CONST$ and PRA_CONST denote the maximum permissible data disclosure risk and public cloud resource allocation cost. In addition, the following observations can be made from the WPP definition of Hybrid-I based on the specifications given above: (i) A data item $R_i \in R$ denotes an attribute of a relation of R, since Hybrid-I uses the relational data model, a vertical data partitioning model and a partial data replication model. (ii) Rep consists of two representations, namely "plaintext" and "bucketization", since Hybrid-I uses a column-level sensitivity model along with bucketization as the security model for public clouds.

[2] `http://hive.apache.org/`

Again, we provide specifications of performance, data disclosure risk and resource allocation cost that aptly capture aspects of the metrics that are relevant to the problem domain modeled by Hybrid-I.

Performance: As stated earlier, Hybrid-I uses the *overall query execution time* of workload Q, denoted as $ORunT(Q, Q_{pub})$, as an indicator of performance. Consequently, the objective function of WPP aims to minimize the overall execution time of a given query workload Q. The execution time of queries in Q over a hybrid cloud, given that queries in Q_{pub} are executed on a public cloud can be represented as follows:

$$Performance(Q, Q_{pub}) = ORunT(Q, Q_{pub}) = \max \begin{cases} \sum\limits_{q \in Q_{pub}} freq(q) \times runT_{pub}(q) \\ \sum\limits_{q \in Q - Q_{pub}} freq(q) \times runT_{priv}(q) \end{cases}$$

Note that $Q_{pub} \subseteq Q$, otherwise it is undefined. Additionally, $freq(q)$ denotes the access frequency of query $q \in Q$ and $runT_x(q)$ denotes the estimated running time of query $q \in Q$ at site x where x is either a public ($x = pub$) or private ($x = priv$) cloud. In practice, Hybrid-I uses the I/O size of the query execution plan selected for processing q at site x as a replacement for the execution time. The running time of a query q can be estimated based on the selected query plan T for site x (x is a public or private cloud) as follows:

$$runT_x(q) = runT_x(T) = \frac{\sum\limits_{\forall\, operator\ \rho \in T} inpSize(\rho) + outSize(\rho)}{w_x},$$

where $inpSize(\rho)$ and $outSize(\rho)$ denote the estimated input and output sizes of an operator $\rho \in T$. Additionally, weight w_x denotes the number of I/O operations that can be performed per unit time at site x. Note that $inpSize(\rho)$ and $outSize(\rho)$ can be computed using statistics accumulated over dataset R for an operator ρ.

Data Disclosure Risk: In Hybrid-I, the risk associated with storing the public side partition of data, namely R_{pub}, using the representations given in Rep, namely plaintext and bucketization, is estimated as follows:

$$Risk(R_{pub}, Rep) = \sum_{R_i \in R_{pub}, s \in Rep} sens(R_i, s),$$

where $sens(R_i, s)$ is the number of sensitive values contained in a data item $R_i \in R_{pub}$, which are stored under the representation, $s \in Rep$, on a public cloud. Finally, the formalization of WPP places a user defined upper bound, $DISC\_CONST$, on the amount of sensitive data that can be disclosed to a public cloud.

Resource Allocation Cost: Hybrid-I estimates the financial cost of utilizing public cloud services as follows:

$$Pricing(R_{pub}, Q_{pub}) = store(R_{pub}) + \sum_{q \in Q_{pub}} freq(q) \times proc(q),$$

where $store(R_{pub})$ represents the monetary cost of storing a subset $R_{pub} \subseteq R$ on a public cloud, $freq(q)$ denotes the access frequency of query $q \in Q$, and $proc(q)$ denotes the monetary cost associated with processing query q on a public cloud. Finally, the formalization of WPP incorporates a user defined parameter, PRA_CONST, which acts as an upper bound on the maximum allowable monetary cost that can be expended on storing and processing data on a public cloud.

5 Conclusions and Future Work

A hybrid cloud is well suited for users who want to balance the efficiency achieved through the distribution of computational workloads with the risk of exposing sensitive information, the monetary costs associated with acquiring public cloud services and the load generated on a private cloud as a result of processing

some part of a workload. In this paper, we identified the criteria that have the greatest impact on the design of an effective hybrid cloud solution and we tabulated the applicability of these criteria to various hybrid cloud deployment models. Then, we formalized the workload partitioning problem as a mechanism for maximizing workload performance using the identified criteria. Finally, we described how existing systems could be derived from the general workload partitioning problem formalization by an instantiation of the appropriate design criteria.

As a part of our future work, we plan to expand on the design criteria we identified in this paper by including factors such as the processing capabilities of a public cloud, which also greatly affect the performance of a hybrid cloud application.

6 Acknowledgements

The work conducted at UT Dallas was partially supported by The Air Force Office of Scientific Research MURI-Grant FA-9550-08-1-0265 and Grant FA9550-12-1-0082, National Institutes of Health Grant 1R01LM009989, National Science Foundation (NSF) Grant Career-CNS-0845803, NSF Grants CNS- 0964350, CNS-1016343, CNS-1111529, and CNS-1228198, and Army Research Office Grant 58345-CS. The work conducted at UC Irvine was supported by the National Science Foundation under Grant No. 1118127.

References

[1] Hybrid Cloud. The NIST Definition of Cloud Computing. *National Institute of Science and Technology, Special Publication, 800-145*, 2011.

[2] K. Zhang, X–y. Zhou, Y. Chen, XF. Wang, and Y. Ruan. Sedic: privacy-aware data intensive computing on hybrid clouds. In *ACM Conference on Computer and Communications Security*, pages 515–526, 2011.

[3] K. Y. Oktay, V. Khadilkar, B. Hore, M. Kantarcioglu, S. Mehrotra, and B. Thuraisingham. Risk-Aware Workload Distribution in Hybrid Clouds. In *IEEE CLOUD*, pages 229–236, 2012.

[4] M. Lev-Ram. Why Zynga loves the hybrid cloud. `http://tech.fortune.cnn.com/2012/04/09/zynga-2/?iid=HP_LN`, 2012.

[5] L. Mearian. EMC's Tucci sees hybrid cloud becoming de facto standard. `http://www.computerworld.com/s/article/9216573/EMC_s_Tucci_sees_hybrid_cloud_becoming_de_facto_standard`, 2011.

[6] Accenture Technology Vision 2011 - The Technology Waves That Are Reshaping the Business Landscape. `http://www.accenture.com/us-en/technology/technology-labs/Pages/insight-accenture-technology-vision-2011.aspx`, 2011.

[7] M. R. Fouad, G. Lebanon, and E. Bertino. ARUBA: A Risk-Utility-Based Algorithm for Data Disclosure. In *Secure Data Management*, pages 32–49, 2008.

[8] S. Trabelsi, V. Salzgeber, M. Bezzi, and G. Montagnon. Data disclosure risk evaluation. In *CRiSIS*, pages 35–72, 2009.

[9] Using Data Encryption. `http://docs.amazonwebservices.com/AmazonS3/latest/dev/index.html?UsingEncryption.html`

[10] J. A. Halderman, S. D. Schoen, N. Heninger, W. Clarkson, W. Paul, J. A. Calandrino, A. J. Feldman, J. Appelbaum, and E. W. Felten. Lest We Remember: Cold Boot Attacks on Encryption Keys. In USENIX Security Symposium, pages 45–60. USENIX Association, 2008.

[11] H. Herodotou. Hadoop Performance Models. Technical Report CS-2011-05, Computer Science Department, Duke University.

[12] H. Hacigümüş, B. R. Iyer, C. Li, and S. Mehrotra. Executing SQL over encrypted data in the database-service-provider model. In *SIGMOD*, pages 216–227, 2002.