

A System for Ensuring Data Integrity in Grid Environments

Austin Gilbert, Ajith Abraham and Marcin Paprzycki
Department of Computer Science, Oklahoma State University, USA
{austirg, aa, marcin}@cs.okstate.edu

Abstract

Data integrity has to become one of the central concerns of large-scale distributed computing systems such as the Grid, whose primary products are the results of computation. In order to maintain the integrity of this data, the system must be resilient to diverse attacks and tampering. The system should also encourage positive influences on its integrity in addition to discouraging or eliminating negative ones. In this paper we develop a model of trust for Grid participants based on the use of reputation systems and associated feedback mechanisms.

Keywords: *Grid security, data integrity, reputation system, distributed trust*

1. Introduction

In these security conscious times the word *integrity* has taken on significance beyond its mere technical meaning. The increased focus on security in almost every branch of Computer Science has spread the word *integrity* into diverse contexts: we have file and database integrity, application integrity and communications integrity, among other uses. Despite this diversity, the underlying meaning of the term has remained relatively independent of its application: *integrity* is the assurance that something is what it propounds to be.

While this is not necessarily the case at present, data *integrity* has to become one of the central concerns of large-scale distributed computing systems such as the (computing) Grid. In order for the Grid to be successful, users must be able to trust the results of Grid computation as much as they trust their desktop applications today. Fostering viable trust in the Grid depends on the integrity of the individual elements within the Grid: files, data, network links and other resources and relationships. The primary concern of this paper is to provide system-level assurance for the Grid. In particular, we focus on *data integrity*: the assurance that all system data is protected from tampering or other malicious attacks. Data coming in and out of the Grid environment must not be modified during transit, storage or processing. Furthermore, we must also be able to trust that the data results of Grid processing are not egregiously errant.

Current approaches to Grid security (typified by the

Globus system [7, 8, 9]) focus almost entirely on the communication/transport aspect of data integrity, employing authentication and encryption technologies to guarantee the safety of data in transit.

While communications guarantees are a vital part of the system's overall integrity, focusing on them alone, to the exclusion of other aspects of the system, is a recipe for disaster. More precisely, we must also be concerned with *who* we are communicating with and *what* is communicated, not only *how*. A large-scale distributed computing environment that doesn't protect result integrity as well as communications is vulnerable to a variety of data and result tampering attacks. There are literally dozens of attack scenarios which would render Grid results useless. It is these scenarios that we must strive to mitigate. The SETI@Home project, perhaps the most well known example of large-scale distributed computing, has already experienced data integrity woes due to unknown and untrusted entities tampering with the computation process [19]. In order to maintain the data integrity of the Grid, we must prevent or mitigate such attacks.

The remainder of the paper is structured as follows: Section 2 introduces data integrity as a function of trust and the use of reputation systems in fostering trust in uncertain environments (e.g. the Internet). Section 3 describes our plan for implementing a reputation system in the Grid environment. Finally, conclusions are drawn in Section 4.

2. Discussion

Computation is the primary function of the Grid. It involves the input data "fed into the system" and the data produced by this computation as the Grid's primary product. From the user's perspective, the processes of the Grid should be so reliable that they are essentially invisible. In order to achieve that level of reliability, we must maintain a high standard of data integrity, and ensure the results of Grid computation are precise, accurate and trustworthy. We must know exactly where results are coming from, and be able to verify this with 100% accuracy. We must know with certainty that data was not modified in transit without detection.

Unfortunately, the problem of data integrity does not beget a straightforward technical solution. The system depends on *trust*, not just technology. A reliable metric of

trust is essential to measuring and maintaining the integrity of data being processed/produced in a Grid environment. If the data is allowed to fall into the wrong hands the probability of its malignant use dramatically increases. Likewise, if input into the system comes from malicious sources, the computational results of the system are likely to reflect that fact, to the detriment of the system's overall integrity. Clearly, our efforts must focus on keeping trusted data and computation from untoward manipulation. This focus is complicated by the observation that in a relatively open environment, such as the Grid, nodes may come and go at any time. Dealing with new and unknown entities presents a difficult trust problem -- a problem of uncertainty. How can we distinguish trusted nodes from untrusted ones when we cannot assume a prior relationship exists? We require a means of measuring and managing trust between nodes in the Grid.

In order to satisfy this requirement, we must look outside Computer Science. Here we find a socioeconomic model of trust that can be adapted to fit our environment: the reputation system.

2.1. Reputation systems

There are three types of reputation systems: positive reputation systems, negative reputation systems and hybrids. A positive reputation system rewards good behavior, in order to encourage a desired outcome. A negative reputation system, in contrast, punishes undesirable behavior. The actors in both types of reputation system start with a neutral reputation. Points are either deducted or added depending on behavior of the actor (in a negative reputation system, points are taken away only as punishment, while in a positive reputation system points are added only as a reward for good behavior). In a hybrid system, both kinds of behavior (positive and negative) are accorded point values. The point distribution of a hybrid system produces broader gradients between desirable and undesirable behavior, making it easier to distinguish between nodes with *good*, *bad*, and *neutral* reputation. >From a sociological perspective, these categories correspond to social contacts we admire, dislike and still others of whom we have no strong opinions.

2.2. The use of hybrid reputation systems

Keser has demonstrated that in markets with high levels of anonymity, the probability of fraudulent activity increases. The introduction of a reputation system effectively counters this tendency [1]. Unfortunately, there remains the problem of exchangeable identities. The easy rotation of identities undermines the benefits of a reputation system [1, 2]. This is true of online markets as well as traditional markets and was also demonstrated by Keser [1].

The solution to this latter problem lies in the use of a weighted hybrid reputation system. More specifically, a hybrid system that is weighted toward rewarding desired behavior ensures that: 1) positive reputations are difficult but not impossible to acquire, and 2) good reputations are relatively easy to diminish since the hybrid system allows fluctuations of a node's reputation in both directions. With this weighted system, "preferred" nodes are much easier to distinguish than in other configurations (there exists a clear distinction between the "good" and the "bad" nodes). This preference exerts a selective pressure on the system's participants. It can be expected that nodes with better reputations will be selected more often than nodes with bad reputation. Yamagishi [3] and Cook [2], show that by making it difficult to obtain good reputations, reputations become a kind of investment, thus encouraging the reuse of entities' identities and discouraging their exchange (which is equivalent to putting one's investment at risk).

In a real Grid economy [4], an investment in one's reputation translates to an actual investment, and accompanying economic success or failure. Nodes with better reputations will commend a higher pay rate than other nodes, just as companies with good reputations in the real world are able to sell their products at higher prices.

2.2.1. E-Bay. The profile system on eBay [14] is analogous to a hybrid reputation system, with buyers and sellers affecting each other's reputations by giving feedback on the interaction experience. Sellers who describe their products accurately, ship on time, etc. and buyers who pay on time, give the correct address, etc. have better reputations than buyers and sellers who abuse the principles of the system. Over time an actor's reputation becomes the embodiment of their actions. If the actor typically generates fair and honest transactions, then a positive reputation will indicate this, and if the actor typically cheats his partners a low or negative reputation would result. When correlated with a history of transactions, an actor's reputation value is a fairly accurate means of revealing trends in his behavior. We will see shortly how such trends becomes the basis for judging the trustworthiness of a potential interaction partner.

2.4. Not as easy as eBay

The positively-weighted hybrid reputation system fits the Grid handedly, however it can't be as simple as eBay's implementation. The eBay system is human driven with optional inputs. To be scalable and effective in the Grid environment, the reputation service must (a) be automatic and (b) mandated. Hence every transaction experience between two peers must be quantified and accounted for by affecting changes in each peer's respective reputations. Obviously, there are many potential approaches for conceptualizing and synthesizing *trust* in the Grid

environment. While in this note we discuss some of the important features of such a system, determining an adequate approach for doing so will be one of the directions of our continuing research.

2.5. Redundancy and feedback

The reputation system alone would likely be inadequate. Reputation systems are driven by input, input which is essentially feedback from other aspects of a system, namely the interaction or transaction mechanisms. The introduction of work redundancies [19] is one means of providing feedback and detecting suspicious return data. The application of redundant result checks and a reputation system taken together will help to ensure the level of data integrity needed in the Grid environment by establishing a system of checks and balances. This system is the basis for trustworthy results produced in the Grid. Without this trust, data integrity cannot exist. Without data integrity, the computational results produced on grids may be inaccurate, rendering them useless.

The primary factor in determining a node's reputation is the quality of the results it produces. This quality is measured/quantified through the use of spot checks and work redundancies. Spot checks will consist of randomly reworking jobs in order to verify the accuracy of the original results. While the reworking of this data can be done by any trusted node excluding the original computer, the system must still be protected from collusion. Detecting and defeating collusion within the system is a topic worthy of its own discussion and falls beyond the scope of this note. A simple approach was outlined in [19] in which work units were redundantly processed and the results checked against each other. Discrepancies in any of the results potentially indicated foul-play. While not perfect, such a solution can substantially improve the overall trust in the system.

A public and visible, feedback-driven reputation system raises a user's confidence in his transactions with system peers and helps steer users around negative experiences by limiting contact with low ranking peers. This is essential in an environment where users must make decisions about other users and the system with a minimum amount of foreknowledge available. The successful application of reputation systems in online auctions leads us to believe that a system with the same underlying precepts would be well-suited to the minimal-knowledge economy of the Grid.

3. Applying reputation systems to the Grid

The Grid is to become the largest existing distributed architecture, therefore the reputation system must be easily distributed and flexible enough to be used by widely varying Grid applications. Our implementation seeks to meet these requirements through the use of X.509

Attribute Certificates to securely distribute and store reputation information.

3.1. Trust synthesized from attributes

In developing the reputation system, we believe that rather than having a single value representing the *trustworthiness of a node*, the reputation should be broken into separate attributes such as but not limited to: speed, accuracy, availability and consistency. Here speed might indicate how quickly a resource is able to complete a task and return results. Accuracy obviously would reflect the historical accuracy of its work. Availability would typically address the nodes frequency of participation within the particular application. Finally, consistency would indicate how consistent is a given node in delivering the result in a promised time-frame. These attributes are presented to exemplify how to break trust into separate *confidences*. Each attribute represents a *confidence*, and each *confidence* represents a characteristic of a node from which trust can be synthesized. There are varying forms of trust. We can *trust* a node to be accurate (this is important for data integrity). We can *trust* a node to complete tasks reliably. We can *trust* nodes to return data quickly, or always in a guaranteed time, so on and so forth. When developing a reputation system each characteristic should be independent of the others to provide flexibility. In such a way as people *trust* physicians for medical advice and stock brokers for financial advice, attributes should be viewed as foundational characteristics used to build particular types of *trust*. Any number of attributes could potentially exist in an application resulting in a varied range in trust types. There could be multiple attributes describing a resource's performance on specific tasks within the application. Using a flexible and extensible basis for synthesizing varying types of trust allows for the greatest flexibility from one application to the next. Flexibility is essential as anything too rigid cannot be easily adopted in a grid environment.

3.2. Simple, lightweight, and distributed

We cannot make too many assumptions about grid participants. Hard drive space may be a luxury for some resources. As such, trust-management data stored by each node should be constrained to a minimal necessary amount. Participating nodes will store their own reputation information encapsulated in a *referral*. The referral will be issued by resources distributing work and the contained attributes will be updated after each interaction. A referral is constructed from an X.509 Attribute Certificate (X.509 AC) where each of the system's attributes is stored in the attribute array of the X.509 AC [18]. The X.509 certificates are then digitally signed by the referrer to provide a chain of accountability

and to prevent modification.

X.509 certificates have limited lifetimes, or periods for which the certificates are valid. Certificate systems must provide a means of revoking valid certificates. A list of revoked certificates, Certificate Revocation List (CRLs), are typically distributed by certificate authorities on a periodic basis. The distribution of CRLs represent unwanted overhead, fortunately CRLs can be avoided in our proposed implementation by limiting the lifetime of each certificate. In the Grid environment, the lifetime of the X.509 certificates should be *short*, meaning that they should not exceed the amount of time expected to complete a task or the amount of time expected to lapse between interactions. We speculate the lifetime of attribute certificates in our implementation should rarely exceed one or two weeks. With such a short lived validity - one week compared to one year on typical identity certificates - the distribution of CRLs are no longer needed. Instead, the issuer of the certificates keeps only a list of valid certificates and all other certificates are considered invalid.

Functionally speaking, the Attribute Certificates will be used in the system much like any other certificate. Our reputation system will utilize the Attribute Certificates in compliment with Identity Certificates provided by the existing infrastructure [8,10]. Where the identity certificates are used to verify the identity of an entity in a highly anonymous environment (such as the Internet), the attribute certificates will be used to determine the trustworthiness of an entity in an uncertain environment (the Grid).

4. Conclusions

In short, the Grid is not just another distributed system, it is a large scale information system. As such, there is an essential need for information integrity within the Grid. This paper briefly introduced why data integrity is important. We believe that achieving data integrity within the grid will require the use of a non-traditional security schema. One such schema for managing trust is the sociological reputation model outlined in this paper in corroboration with work redundancies described in [19]. We concede that more research will be required to validate this approach in the grid environment, but again point to indications from studies in other fields that using reputation systems has a positive effect on market-like interactions: [14], [3], [1], [2]. To this end, a simulation of the model is being developed to study interactions of grid elements utilizing a reputation system to manage trust. We also hope to present a deeper thesis of the technical arguments in the near future.

5. References

- [1] C. Keser, "Experimental Games For the Design of Reputation Management Systems", *IBM Systems Journal*, 42(3), 498-506, 2003.
- [2] T. Yamagishi, "The Role of Reputation In Open and Closed Societies: An Experimental Study of Internet Auctioning", in Proceedings, First Interdisciplinary Symposium On Online Reputation Mechanisms, Cambridge: Massachusetts, U.S.A, April 2003.
- [3] K.S. Cook, E.R.W. Rice, A. Gerbasi, "Commitment and Exchange: The Emergence of Trust Networks Under Uncertainty", In proceedings, Formal and Informal Cooperation Workshop, Collegium Budapest, November 2002.
- [4] R. Buyya, D. Abramson, and J. Giddy, "A Case for Economy Grid Architecture for Service Oriented Grid Computing", 10th IEEE International Heterogeneous Computing Workshop (HCW 2001), In conjunction with IPDPS 2001, San Francisco, California, U.S.A, April 2001.
- [5] A. Abdul-Rahman and S. Hailes, "Using Recommendations For Managing Trust In Distributed Systems", In Proceedings 1999 International Symposium on Database, Web, and Cooperative Systems (DWACOS'99), Baden-Baden, Germany, August 1999.
- [6] K. Aberer and Z. Despotovic, "Managing Trust in a Peer-2-Peer Information System", In proceedings, 2001 ACM CIKM International Conference on Information and Knowledge Management, 2001.
- [7] I. Foster, N.T. Karonis, C. Kesselman, G. Koenig, and S. Tuecke, "A Secure Communications Infrastructure for High-Performance Distributed Computing", In proceedings, 6th International Symposium on High Performance Distributed Computing (HPDC '97), Portland, Oregon, U.S.A., August 1997.
- [8] I. Foster, C. Kesselman, G. Tsudik, and S. Tuecke, "A Security Architecture for Computational Grids", In proceedings of the 5th ACM Conference on Computer and Communications Security Conference, 82-92, November 1998.
- [9] V. Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, S. Tuecke, "Security for Grid Services", ANL/MCS-P1024-0203, February 2003. In proceedings, 12th IEEE International Symposium on High Performance Distributed Computing (HPDC'03), Seattle, Washington, U.S.A., June 2003.

[10] L. Pearlman, V. Welch, I. Foster, C. Kesselman, and S. Tuecke, "A Community Authorization Service for Group Collaboration", In proceedings, 3rd International Workshop on Policies for Distributed Systems and Networks (POLICY '02), Monterey, California, U.S.A. June 2002.

[11] K. Keahey, and V. Welch, "Fine-Grain Authorization for Resource Management in the Grid Environment", In proceedings, Grid Computing 2002 (GRID 2002), Third International Workshop, Springer, *Lecture Notes in Computer Science: Security and Policy Management*, 2536, Baltimore, MD, U.S.A., November 2002.

[12] M. Lorch and Dennis Kafura, "Supporting Secure Ad-hoc User Collaboration in Grid Environments", Available: <http://people.cs.vt.edu/~kafura/RecentPapers/>

[13] M. Blaze, J. Feigenbaum, J. Ioannidis, and A.D. Keromytis, "The Role of Trust Management in Distributed Systems Security", *Secure Internet Programming: Issues in Distributed and Mobile Objects Systems*, Springer-Verlag Lecture Notes in Computer Science *State-of-the-Art* series, 185-210, Berlin 1999.

[14] D. Houser and J. Wooders, "Reputation in Auctions: Theory and Evidence from eBay", Available: <http://databases.si.umich.edu/reputations/bib/bib.html>

[15] A. Zieger, "Grid Security: State of the Art", available: <http://www-106.ibm.com/developerworks/library/gr-security.html> (2003)

[16] A. Abdul-Rahman and S. Hailes, "A Distributed Trust Model", In proceedings ACM New Security Paradigms Workshop '97, Cumbria, UK, September 1997.

[17] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara, "Reputation Systems", *Communications of the ACM*, **43**(12), 45-48, 2000.

[18] S. Farrell and R. Housley. RFC 3281 Standard: An Internet Attribute Certificate. (2002)

[19] D. Molnar, "The Seti@Home Problem", *ACM Crossroads*, September 2000, available: <http://www.acm.org/crossroads/columns/onpatrol/september2000.html>