GRID COMPUTING IN EXTREME SITUATIONS: REDUCING COSTS AND CREATING RESILIENCE FOR IT INFRASTRUCTURES

Oliver Hinz, Chair of Business Administration, esp. Electronic Commerce, J.W. Goethe University Frankfurt, Mertonstrasse 17, 60325 Frankfurt, ohinz@wiwi.uni-frankfurt.de

Michael Schwind, Chair of Business Information Systems and Operations Research, Technical University Kaiserslautern, Erwin Schrödinger Strasse, Geb. 42, 67653 Kaiserslautern, schwind@wiwi.uni-kl.de

Roman Beck, Chair of Business Administration, esp. Information Systems, J.W. Goethe University Frankfurt, Mertonstrasse 17, 60325 Frankfurt, rbeck@wiwi.uni-frankfurt.de

Abstract

Recent turbulences in financial markets are not only a challenge for the actors in the front offices of the related institutions, but also represent a serious challenge for the IT departments in the back offices of banks etc. We present a simulation model that shows how Grid computing increases the resilience and quality-of-service of IT infrastructure in departmentalized enterprises in the presence of shocks. Grid computing also reduces the costs deriving from the cancellation of jobs in times with a high volatility of computational load. The model can be used to find the appropriate type of IT infrastructure for different financial service institutions. Our simulations' findings are also likely to encourage the introduction of Grid computing for related business branches and applications.

Keywords: Grid computing, resilience, IT infrastructure, exogenous shocks, quality-of-service

1 INTRODUCTION

Increasing competition on global markets leads to high pressure on enterprises and consequently requires further restructuring and automation of IT-related business processes such as in the financial services industry. In addition to the stiffening competition, banks have to cope with new legal regulations such as Basel II and customer needs that are changing in the direction of highly customized on-demand financial products. Finally, the closely woven and interconnected international financial markets react extremely sensitively to any relevant information such as national financial turbulences, market crashes or bubble bursts that lead to highly volatile markets and high trading volumes that are difficult to forecast.

Such extreme events in financial markets not only cause turbulences for the related actors in the market institutions but also have a significant impact on the critical IT infra-structure of these institutions. If the trading volume of the markets rises to a tenfold (or even more) of the regular level in extreme market situations, it is very likely that IT infrastructure is not capable of dealing with this load. A failure or slow down of computing services can cause serious damage to institutions in the finance industry especially in critical situations, e.g. if trades, transactions, and settlements of financial instruments and products cannot be guaranteed just in time. However, maintaining an infrastructure that can deal with such rare events is also very expensive for the financial institutions. Our scenario deals with an enterprise of the financial industry (e.g. a bank) with a separate computing department for each organizational unit. In this paper we show for the case of such a departmentalized bank that the introduction of Grid computing can significantly increase IT architecture *resilience* and thereby *quality-of-service* (*QoS*) with lower costs for the cancellation of requested computing jobs even in extreme situations, e.g., those caused by financial crashes etc.

In order to understand the *vulnerability to exogenous shocks* of a widespread IT organization structure better, we designed a simulation system that is able to model a very common organization scheme of IT infrastructure in big European financial institutions. By feeding load profiles that are typical for extreme situations into the departmentalized IT infrastructure, we compare a *Grid* and a *Non-Grid* solution with respect to their *resilience* and *quality-of-service* (*QoS*). This is only the first step in our research process; tests for real world load-profiles will follow to substantiate our simulation findings.

The remainder of the paper is structured as follows: Section 2 provides a categorization of Grid computing in the light of management implications, followed by a detailed discussion of the importance of resilience for industries which depend on the availability of up-to-date information in section 3. Section 4 will introduce the developed and applied simulation model for Grid computing in extreme cases, while we will discuss the findings and conclude the paper in section 5.

2 MANAGEMENT PERSPECTIVES OF GRID COMPUTING

To our knowledge the management perspectives of Grid computing are not widely discussed in recent literature yet. Especially for aspects of system resilience, there is a research gap. For this reason, we try to classify the few existing papers using a classical scheme of managerial literature.

We identify three major management perspectives for Grid computing technology following the classical differentiation scheme with respect to the short, middle, and long term planning horizon of an enterprise (Ansoff, 1965):

2.1 Strategic management perspective of Grid computing

- *Globalization of computing:* The concept of computing around the world is gaining increasingly in importance. The expansion of the enterprises' core business into new geographical regions (e.g. China) plays an important role in the business strategy of many international companies. and Global Grid computing can reduce investments into IT infrastructure in the case of such an expansion strategy, because IT resources located in other countries can be co-used (2002). This strategy is often called 'follow the sun' because the time zones in the different geographical regions lead to resource load profiles that follow the rhythm of day and night.
- *Sustainable resource management:* Ecological factors are gaining increasingly in importance for the IT industry and companies with big IT departments. 'Green computing' aims at the effective use of computing resources to preserve natural resources. The trend has attracted the interest of governmental institutions as well as industrial companies (GreenerComputing, 2007; GreenGrid, 2007). Several experiments have shown that a suitable organization of Grid computing infrastructure can lead to a significant reduction in power consumption (Patel et al., 2002).
- *Resilience of computing:* Another strategic view of Grid computing includes reaction to exogenous shocks. The extreme increase in IT resource load due to an unexpected event like a crash in the financial markets that leads to a huge demand temporary for computational power is an example of such an exogenous shock (Huang & Bhatti, 2004). Most recent literature includes some aspects of network resilience (Castain & Squyres, 2007) and robust allocation mechanisms for computational tasks on Grid systems (Choon Lee & Zomaya, 2007) Resilience to exogenous shocks in an entire Grid infrastructure has not been discussed in much detail yet.

2.2 Tactical management perspectives of Grid computing

- *Cost reduction:* A major target that is expected to be achievable by introducing Grid computing technology in the IT service infrastructure of companies which have an increased demand for computational power is the significant reduction of resource costs (Skillicorn, 2002). The reorganisation of IT systems by introducing Grid technology in order to improve the usage of the existing resources can be considered as a tactical variant of cost reduction. However, exact figures about the true dimension of possible cost reduction are not known in the literature yet (Cheliotis et al., 2004).

2.3 Operative management perspectives of Grid computing:

- *Failure risk reduction and quality of service:* The application of parallel processing in distributed computer systems, as is the case in Grid system architectures, makes it possible to reduce the risk of failure of the entire system and helps to increase the reliability and robustness of service provision (Baker et al., 2002; Czajkowski et al., 2001). Consequently Grid systems are usually able to offer a higher level of *QoS* for IT service provisioning processes while employing fewer resource capacities as a result of the increased system reliability and robustness (Schwind et al., 2007). This rise in *QoS* is mainly of interest from the operative point of view. However, as for the case of IT cost reduction by the application of Grid systems, exact figures about the achievable level of improvement are not yet known for industrial real world applications.

- *Scheduling and load balancing:* Another advantage of Grid systems from the operative point of view is their ability to schedule incoming computing tasks on the pooled IT resources and to achieve a suitable load balancing. Economically-oriented Grid scheduling systems in particular have turned out to be a promising approach for load balancing in distributed computer systems (Eymann et al., 2003; Schwind et al., 2006).

While regarding the positive impacts of Grid computing, one should not forget about some economic downside risk of Grid computing. Though there is a legitimate opportunity of cost reduction, there can also be additional hidden costs of Grid computing. These costs might result from deployment, update, and version management effort, resulting from the distributed hardware infrastructure of a Grid system (Afgan & Bangalore, 2007). Additionally, sufficient network capacity is required for a reliable working Grid infrastructure (Huang & Bhatti, 2004).

3 RESILIENCE IN GRID COMPUTING

Grid architectures can be implemented not only to optimize existing business processes by faster calculation or availability of data but also to improve the resilience of the enterprise's IT infrastructure to hardware failure or unforeseen peak loads. As mentioned already in the previous section, computing resilience is an important factor from a strategic IT management perspective.

For instance, a German headquarters can use the office infrastructure of their subsidiary company in the US to augment processing power and to run applications during the American off-peak hours. There is significant potential for peak demand clipping within organizations that straddle multiple time zones (Skillicorn, 2002). This flexibility and scalability of capacity allows it to provide computing capacity to meet average demand, taking advantage of virtualized resources, to meet unexpected surges of resource requirements, and improve the utilization of existing IT assets. In addition, departmentalized enterprises can take full advantage of this to use underutilized computing resources to serve as backup and recovery systems for improved operational resilience and reduced infrastructure investment requirements.

According to Xie et al. (2005), resilience is the ability of IT infrastructure to guarantee a certain level of service in the presence of sudden imponderabilities such as natural disasters, failures due to operational errors, attacks on the IT, or unpredictably long delay paths. From a management perspective, erratic but extreme volatilities of usage can be added to the potential threats an IT infrastructure has to cope with. Consequently, resilience in this paper is defined as the IT system's ability to provide a certain predefined *QoS* even if an unusual high but legitimate traffic load occurs (Menasce & Casalicchio, 2004a; Menasce, 2004). In this context the ability to measure the *QoS* in Grid systems plays a role of high importance (Colling et al., 2007; Menasce & Casalicchio, 2004b).

More traditional approaches for pooled computer resources address the topic of resilience under the term "high availability" *HA* especially in connection with earlier cluster computing applications (Gray & Siewiorek, 1991). The *HA* concept defines classes for the time ratio of a computer system's *uptime*[1] to its *uptime* plus *downtime* and such provides a definition for *QoS* standard. In contrast to our approach which ensures *QoS* by the pooling of computer resources (Grid) that may be heterogeneous or not, *HA* is mainly concerned with the design of fault tolerant (resilient) hardware architectures by creating redundant structures (backup solutions) in homogeneous computer environment.

IT resilience is especially important if the critical infrastructure is supporting complex adaptive systems (Holland, 1968) such as capital and stock markets with their low-probability/high-consequence events. In "millisecond" industries such as the financial services industry the IT

[1] Uptime is the time the computer system is available for providing services requested by the users; downtime defines the time the system is not available.

infrastructure requires significant attention to resilience. Operating at high speed, information-based industries require Grid architecture-like IT infrastructure to strengthen resilience, diversity, and redundancy. Thus acknowledging the role of resilience in critical infrastructure can reduce operational risk in extreme situations.

4 MODEL AND SIMULATION FOR GRID COMPUTING IN EXTREME SITUATIONS

4.1 Simulation Model

To assess the impact of exogenous shocks like political or financial crises on the IT infrastructure of highly departmentalized enterprises, we conduct a simulation study and compare two different IT infrastructures: Firstly, we look at an IT infrastructure where every department has exclusive access to dedicated servers. This is the traditional way access to IT resources is organized in the financial services industry today. In the second case, the IT resources are pooled by means of virtualization technique. We call this IT infrastructure a Grid. In both cases a single IT controller gathers the demand requests of the consuming departments and optimizes the number of servers needed.

The enterprise thus consists of two types of players: First, the enterprise consists of n departments that demand IT resources. For the sake of simplicity, all departments i demand d_i of an uniform IT resource given by the normal distribution $N(\mu_i, \sigma_i^2)$. Beside the demand d_i, these IT consuming departments i vary in their preferences by facing cancellation costs c_i. Departments with urgent or important jobs suffer from cancellation more than departments with low priority jobs and therefore demand a higher level of QoS. The required level of QoS_i is thus a function of c_i and will be explained later.

Every department submits its demand function to the second type of agent, the IT controlling agent. There is only one instance of IT controlling aggregating the demand. The IT controlling agent calculates the amount of computational power that is needed to ensure the QoS required for all departments. Every additional server unit costs s and is assumed to be constant and exogenously given, since a single enterprise does not have an influence on the market price of servers. Every server delivers m units of uniform IT resource.

The process of a simulated period can be described as follows: Simulation time is discrete (e.g. every time period equals 1 month) and a fixed number of processes are executed for all agents in every simulated period. At the beginning of the simulation, every department calculates the optimal level of quality of services given by the following indifference equation:

$$\left(1 - cdf_{\mu_i,\sigma_i}(d_i)\right) \cdot c_i = d_i \cdot \left(\frac{s}{m}\right) \Leftrightarrow \left(1 - \tfrac{1}{2}\left(1 + erf_{\mu_i,\sigma_i} \frac{d_i - \mu_i}{\sigma_i \sqrt{2}}\right)\right) \cdot c_i = d_i \cdot \left(\frac{s}{m}\right)$$

where d_i is the demand at which department i becomes indifferent between paying the expected cancellation costs and paying for d_i units of uniform IT resource which cost the server costs s divided by the computational power m per server. The rationale behind this equation is that a risk-neutral enterprise is indifferent between paying the cancellation costs and paying for the server costs required for doing the job. By solving this indifference equation, the department can calculate the cut-off demand d_i and can then calculate the optimal level of quality of service by setting d_i into the cumulative distribution function:

$$QoS_i = cdf_{\mu_i,\sigma_i}(d_i) = \tfrac{1}{2}\left(1 + erf_{\mu_i,\sigma_i}\frac{d_i - \mu_i}{\sigma_i\sqrt{2}}\right)$$

After the calculation of the optimal service level, each department submits its demand function and the level of *QoS* demanded to the IT controlling agent. The IT controlling agent thereafter determines the optimal number of servers given the demand and required *QoS* by the following decision rules:

For the Non-Grid-Architecture, the IT controlling agent treats all requests separately and extrapolates the cost of the service for every single department. This is the traditional business practice: Department *i* needs a service and therefore orders dedicated servers for the fulfilment of this service. The monthly costs are kept constant over time.

The IT controlling agent calculates the optimal numbers of dedicated servers for every single department based on its demand request (demand function and level of *QoS*). This can easily be done by calculating the inverse cumulative distribution function of the normal distribution and rounding up, since the number of servers has to be whole-number. The optimal number of servers s_i for a single department *i* is then given by:

$$s_i = \left\lceil \frac{1}{m} \cdot cdf_{\mu_i,\sigma_i}^{-1}(QoS_i) \right\rceil = \left\lceil \frac{1}{m} \cdot \left(\mu_i + \sigma_i\sqrt{2} \cdot erf_{\mu_i,\sigma_i}^{-1}\left(2QoS_i - 1\right)\right) \right\rceil$$

Thereby, the optimal number of servers for the entire bank is given by:

$$\hat{s} = \sum_{i=1}^{n} \left\lceil \frac{1}{m} \cdot \left(\mu_i + \sigma_i\sqrt{2} \cdot erf_{\mu_i,\sigma_i}^{-1}\left(2QoS_i - 1\right)\right) \right\rceil$$

For Grid architectures the IT controlling agent aggregates the demand and allocates enough computational power for the pool of departments as a whole. By aggregating the demand, the variation of demand may even out. In probability theory, if *X* and *Y* are independent random variables that are normally distributed, then *X* + *Y* are also normally distributed. Therefore, the total demand *TD* is given by:

$$TD = \sum_{i=1}^{n} D_i \sim N\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^{\,2}\right)$$

The optimal number of servers for the bank as system is then given by the sum of servers for distinct classes of service levels. Since the demanded *QoS* is heterogeneous, the optimal number of servers should be calculated by building classes of homogeneous *QoS*-classes, summing up the number of servers for these classes:

$$\hat{s} = \sum_{i=0}^{1} s_{class} \quad \text{with}$$

$$\hat{s}_{class} = \left\lceil \frac{1}{m} \cdot cdf_{a,b}^{-1}(QoS_{class}) \right\rceil = \left\lceil \frac{1}{m} \cdot \left(a + b\sqrt{2} \cdot erf_{a,b}^{-1}\left(2QoS_{class} - 1\right)\right) \right\rceil =$$

$$\left\lceil \frac{1}{m} \cdot \left(\sum_{i=1}^{n}\mu_i + \sqrt{\sum_{i=1}^{n}\sigma_i^2}\,\sqrt{2} \cdot erf_{a,b}^{-1}\left(2QoS_{class} - 1\right)\right) \right\rceil$$

with

$$a=\sum_{i=1}^{n}\mu_i;\ b=\sqrt{\sum_{i=1}^{n}\sigma_i^2}$$

After the calculation of the optimal number of servers required and their acquisition, the simulation moves into the second phase where in every time step the actual demand is determined. The actual demand is rolled from the given normal distribution for every department. If the demand can not be fulfilled due to a lack of computational power, the department suffers as a result of its cancellation costs. We calculate the total costs for the entire enterprise, which are given by cancellation costs per department and costs for server maintenance. Overall, the enterprise needs to find the optimal number of servers which should be as low as possible but should also avoid cancellation costs.

However, the predicted demand can vary due to some unexpected shock: In economics a shock is defined as an unexpected or unpredictable event that affects an economy, either positively or negatively. We borrow these constructs from economics and model extreme situations by applying exogenous shocks, meaning events that affect the IT infrastructure negatively, occur rarely and can thus not be taken into account ex ante. In the financial services industry this might happen due to a financial crisis or severe political crises.

We model a shock as an unexpected increase in demand for IT resources. This is a common practice in macroeconomics (e.g. Hickman & Klein, 1984). The risk of a shock can be expressed as the probability density of the consequences. In economics risk has therefore two dimensions: The occurrence probability *p* and the weight of the consequences *w* (Zweifel & Eisen, 2002, p. 34*)*.

Though the departments assess their demand as $d_i \sim N(\mu_i, \sigma_i^2)$, their actual demand is given by $d_i \sim N(\mu_i * x, \sigma_i^2)$ with *x* representing additional demand excited by the shock. As suggested by economic theory, the shock is not expected and should thus not be taken into account when departments report their expected demand to the IT-controlling agent.

The variable *x* consists of two parts, $x=w*z$: Firstly, *w* models the weight of the consequences and due to the multiplicative composition the consequences in terms of additional demand are relative to μ_i. If a shock occurs, the demand is then e.g. increased by 50% for the department. Secondly, *z* represents the stochastic part and depends on the occurrence probability *o* of the shock. *z* is a binomial variable given by the following rule: *z=1, if random number* ≤ *o and z=0 otherwise.*

Normally, one would expect different departments to be affected to different extents by the shock. However, we assume for the sake of simplicity that a shock influences all departments in the same manner. Nevertheless, the absolute influence of the shock varies due to the relative combination with the demand μ_i. Note that the actual demand function is not normally distributed anymore since it is biased with this demand shock.

As dependent variable we observe the quality of service and the total costs and compare two different IT-infrastructures: Firstly, we run the simulation for a departmentalized enterprise with dedicated servers as benchmark, and secondly we evaluate the same scenario for an enterprise that pools its resources and applies a Grid IT-infrastructure by means of resource virtualization.

Schwind et al. (2007) show that Grid technology can drive down costs and increase the *QoS* delivered in the absence of shocks. We also expect this advantage to be robust in the presence of shocks and thus hypothesize:

> ***H1a: Grid technology increases the QoS delivered even in the presence of exogenous shocks.***
> ***H1b: Grid technology drives down the total costs even in the presence of exogenous shocks.***

These two hypotheses are not very surprising but we expect that a "gridified" IT infrastructure also increases resilience. We therefore introduce the following two hypotheses that have not been tested quantitatively as far as we know:

> ***H2a: Grid technology increases resilience in terms of QoS in the presence of exogenous shocks.***
> ***H2b: Grid technology increases resilience in terms of maintenance and cancellation costs in the presence of exogenous shocks.***

Our simulation is based on the following assumptions:

- All departments demand a homogeneous, arbitrary, and uniform IT resource. The demand d_i is exogenous.
- Additional IT resources can easily be supplied by acquisition of additional servers.
- The price s for additional servers is constant and exogenously given.
- Departments can evaluate their demand a priori accurately, except the additional demand that can be excited by some exogenous shock.
- A shock has the same relative impact on all departments in the enterprise.

Since we are not using units, results can not be interpreted on an absolute basis. As we are only interested in the impact of the IT infrastructure type total costs, *QoS* and the resilience of the system, we can use these figures to compare on a relative basis.

4.2 Simulation Parameters and Results

We developed the simulation based on the second model described from scratch in c# under .net and use the following initial parameters: We look at an enterprise with 25 departments and thus set n=25. The costs for a server are set to s=10,000 and every server produces m=100 units of uniform IT resources. The average demand for each department is drawn from N~(1000,250) and the standard deviation for this demand is drawn from N~(100,25). Overall, we look at 365 time steps per scenario.

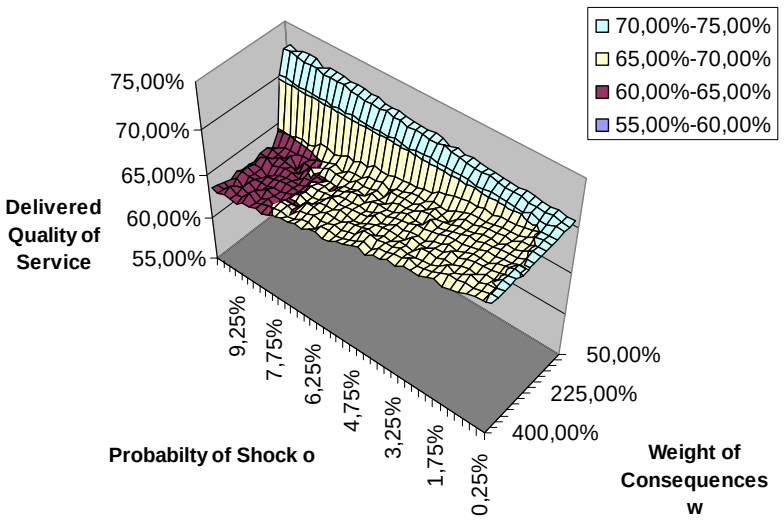

Figure 1. Delivered Quality of Service as Function of o and w for a Non-Grid Architecture

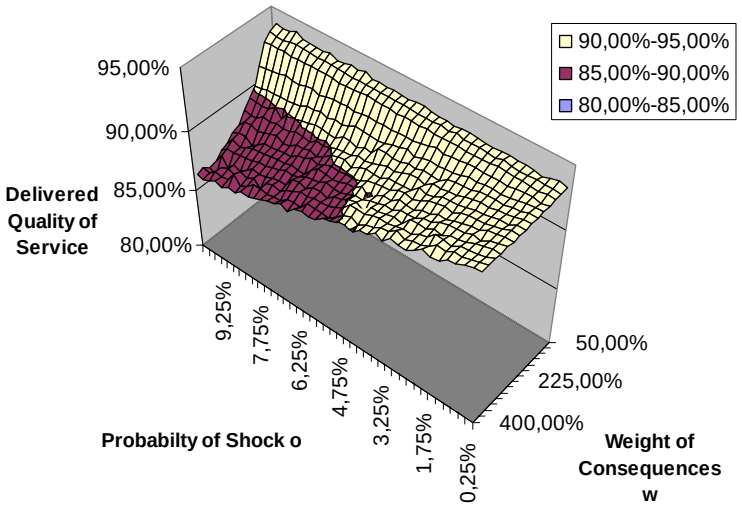

Figure 2. Delivered Quality of Service as Function of o and w for a Grid-Architecture

We vary the probability of occurrence o from 0.25% to 10% and the weight of consequences w from 50% to 400%, w=200% would hence mean that the demand has doubled due to the shock. We also vary the cancellation costs c_i for every department which is drawn from $N\sim(\mu_{cancellationCosts}, \sigma^2_{cancellationCosts})$. $\mu_{cancellationCosts}$ runs from 100,000 to 500,000 and $\sigma_{cancellationCosts}$ from 1 to 200,000.

As depicted in Figure 1 the *QoS* delivered in Non-Grid environments drops from around 75% with increasing probability for shock occurrence o and increasing weight of consequences w since the additional demand was not taken into account. It is also obvious that the decrease in *QoS* is very steep even with small increases in w.

Figure 2 shows the same setting for an enterprise that pools its resources. The *QoS* is on average 91.02% whereas the non-Grid-architecture delivers only an average *QoS* of 67.69%. We conduct an ANOVA (Analysis of Variance) to test hypothesis H1a and find support on the 1%-level ($p<0.01$). Therefore we conclude that the virtualization enabled by Grid technology increases the absolute level of *QoS* even when the system suffers from stochastic shocks.

Surprisingly, this increase in *QoS* due to virtualization comes with lower costs: we use the total costs, which are the sum of maintenance costs and cancellation costs, as dependent variable, normalize it to 1 and test H1b. The ANOVA shows that the total costs using the Grid infrastructure are only 67.02% of the total costs of the equivalent non-Grid-infrastructure and is hence significantly ($p < 0.01$) lower. We thus find support for H1b and conclude that Grid technology leads to higher *QoS* and simultaneously to lower total costs.

Figure 2 also demonstrates that the loss in *QoS* is not as steep with increasing w and/or o as it is in Figure 1. To make this effect more visible we pick a typical scenario and fix w=2.25, $\mu_{cancellationCosts}$=500,000 and $\sigma_{cancellationCosts}$=200,000 and calculate the marginal loss of *QoS* with increasing o. This scenario is depicted in Figure 3. The loss of quality of service per additional point of shock probability is significantly ($p<0.05$) higher with non-Grid-architectures than with Grid architecture.

To test hypotheses 2a and 2b rigorously, we run a linear regression for both dependent variables and compare the coefficients of o and w across the two scenarios (non-Grid vs. Grid).

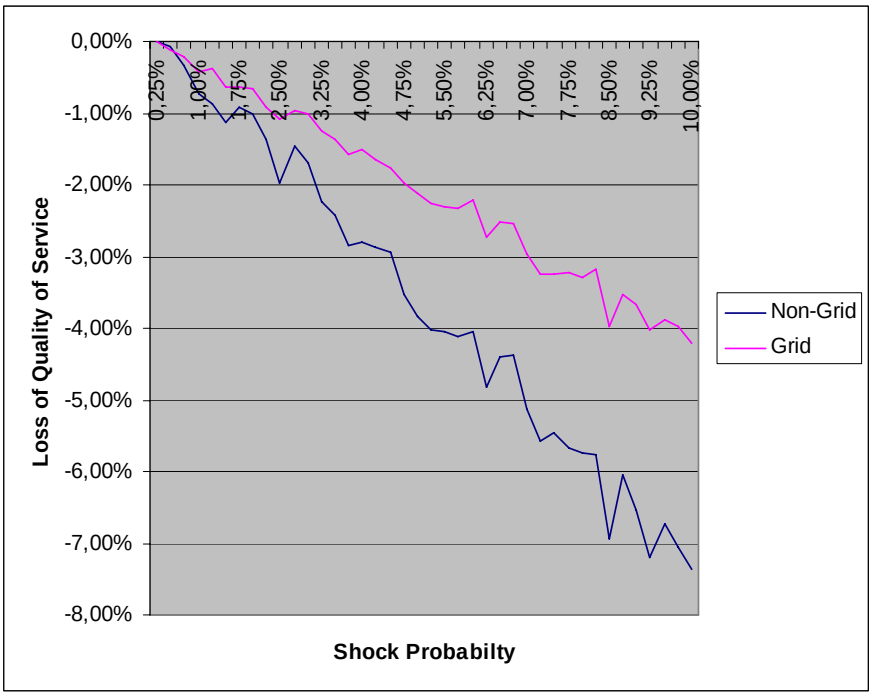

Figure 3. Marginal Loss of QoS with increasing o

Table 1 reveals some interesting insights: Firstly, *QoS* in Grids is typically higher which is depicted by the higher constant. Secondly, higher cancellation costs increase *QoS* for both architectures, since it is cheaper to buy new servers than to drop jobs unfinished when cancellation costs are high. Thus, $\mu_{cancellationCosts}$ has a positive sign. However, the influence of the mean cancellation costs is significantly higher with non-Grid-architecture than with Grid architecture.

Scenario *Dependent Variable*	*Non-Grid* ***QoS***	*Grid* ***QoS***	*Non-Grid* ***Total Costs***	*Grid* ***Total Costs***
Constant	0.354^{***}	0.765^{***}	0.729^{***}	0.506^{***}
$\mu_{CancellationCosts}$	$1.319e\text{-}006^{***}$	$0.6654e\text{-}006^{***}$	$5.781e\text{-}007^{***}$	$2.556e\text{-}007^{***}$
$\sigma_{CancellationCosts}$	$-3.163e\text{-}007^{***}$	$-1.336e\text{-}007^{***}$	$-0.172e\text{-}006^{***}$	$1.534e\text{-}008^{***}$
Shock Probability o	-0.0121^{***}	-0.0118^{***}	0.11^{***}	0.07^{***}
Weight of Consequences w	-0.0051^{***}	-0.0040^{***}	0.23^{***}	0.022^{***}
R^2	79.7%	55.5%	67.1%	75.3%

Table 1. Regression Model with QoS and TotalCosts as Dependent Var. (n=9,000 each), $^{***}p<0.01$

Higher variance in cancellation costs given by $\sigma_{cancellationCosts}$ usually leads to a loss in *QoS*. This is quite intuitive and not surprising. Again, we recognize that the Grid architecture is more robust to a higher variance than the non-Grid-architecture.

Both parameters describing extreme situations, *o* and *w*, have a negative influence on the *QoS* delivered, as expected. However, we again recognize by the smaller coefficients that the Grid-Architecture is more resilient against these shocks. Grid is especially robust in scenarios where the weight of the consequences is high. The defect of departmentalized IT architectures with dedicated servers already became evident in the chart depicted in Figure 1. Hypothesis 2a cannot thus be rejected, and we conclude that Grids are more resilient in terms of *QoS* than equivalent traditional non-Grid-architectures without virtualization.

All these findings equally hold for total costs as dependent variable. Grid technology drives down costs significantly, the architecture is more robust against changes in cancellation costs and finally shocks do not have such a severe influence on costs. Hypothesis 2b cannot be rejected.

5 CONCLUSION

We presented a simulation model that allows us to investigate the consequences of extreme events (e.g. market crashes) for the IT infrastructure of an enterprise in the financial industry that is organized in a departmentalized structure. In a first step our model calculates the optimal size of IT infrastructure (servers) in the enterprise's departments, both for a Grid and a non-Grid-architecture. After this step, the impact of extreme events (e.g. load peaks in the case of a financial crash) on this type of resource allocation is simulated for the two approaches by varying the probability of exogenous shocks and the weight of their consequences. As a result, it turned out that the Grid architecture is not only able to maintain a higher quality-of-service level in the presence of such exogenous shocks compared with a non-Grid-solution, but is also able to reduce the cost for jobs that are cancelled due to the high load situation in the computing system. We were also able to demonstrate that a Grid-solution is far less sensitive to the impact of extreme events than a non-Grid-system, both in terms of the shocks' probability and their consequences. Our simulation results suggest the introduction of Grid computing into business applications due to a significant cost reduction and increased system resilience. If we are able to prove the relevance of our simulation results in real world settings, together with our industrial partners in the financial services industry, this will be the first step into the direction of trading Grid compute resources as a flexible intra-enterprise utility. The next step should be the inter-enterprise exchange of computational resources. This could help to further increase resilience to extreme events, especially if the exchange of compute resources is established between enterprises that belong to different industry sectors. At the moment, however, security concerns prevent such a globalization of Grid computing, especially in the financial industry sector.

References

Ansoff, I. (1965) Corporate strategy. McGraw-Hill, New York.

Afgan, E.; Bangalore, P., (2007) Computation Cost in Grid Computing Environments; In Proceedings of the First Int. Workshop on the Economics of Software and Computation (ESC '07), Minneapolis, Minnesota, USA, pp. 20-26.

Baker, M., Buyya, R. and Laforenza, D. (2002) Grids and grid technologies for wide-area distributed computing. Software: Practice and Experience 32, 1437-1466.

Castain, R. and Squyres, J. (2007) Creating a transparent, distributed, and resilient computing environment: The openrte project. Journal of Supercomputing 42, 107-123.

Cheliotis, G., Kenyon, C. and Buyya, R. (2004) Grid economics: 10 lessons from finance. In Peer-to-peer computing: Evolution of a disruptive technology (Subramanian, R. and Goodman, B., Eds.), Idea Group Publisher, Hershey, PA.

Choon Lee, Y. and Zomaya, A. Y. (2007) Practical scheduling of bag-of-tasks applications on grids with dynamic resilience. IEEE Transactions on Computers 56 (6), 815-825.
Colling, D., Ferrari, T., Hassoun, Y., Huang, C., Kotsokalis, C., Mcgough, S., Patel, Y., Ronchieri, E. and Tsanakas, P. (2007) On quality of service support for grid computing. In Proceedings of the 2nd International Workshop on Distributed Cooperative Laboratories and Instrumenting the GRID (INGRID 2007).
Czajkowski, K., Fitzgerald, S., Foster, I. and Kesselman, C. (2001) Grid information services for distributed resource sharing. In Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), p 181, IEEE Press, Los Alamitos, CA, USA.
Eymann, T., Reinicke, M., Ardaiz, O., Artigas, P., Freitag, F., Messeguer, R., Navarro, L. and Royo, D. (2003) Decentralized vs. Centralized economic coordination of resource allocation in grids. In 1. European Across Grids Conference.
Gray, J. and Siewiorek, D. P. (1991) High availability computer systems. Computer 24 (9), 39-48.
Greenercomputing (2007) Greener computing.
Greengrid (2007) The green grid.
Hickman, B. G. and Klein, L. R. (1984) Wage-price behavior in the national models of project. American Economic Review,
Holland, J. H. (1968) Hierarchical descriptions, universal spaces and adaptive systems. Engineering, College of, Ann Arbor.
Huang, Y. and Bhatti, S. N. (2004) Decentralized resilient grid resource management overlay networks. In IEEE International Conference on Services Computing pp 372-379, IEEE Computer Society, Shanghai, China.
Menasce, D. A. (2004) Mapping service-level agreements in distributed applications. IEEE Internet Computing 8 (4), 100-102.
Menasce, D. A. and Casalicchio, E. (2004a) Qos in grid computing. IEEE Internet Computing 8 (4), 85-87.
Menasce, D. A. and Casalicchio, E. (2004b) Quality of service aspects and metrics in grid computing. In Computer Measurement Group Conference 2004, Las Vegas, NV.
Patel, C., Sharma, R., Bash, C. and Graupner, S. (2002) Energy aware grid: Global workload placement based on energy efficiency. HP Laboratories, Palo Alto.
Schwind, M., Gujo, O. and Stockheim, T. (2006) Dynamic resource prices in a combinatorial grid system. In Conference on E-Commerce Technology (CEC'06), IEEE Press, San Francisco, US.
Schwind, M., Hinz, O. and Beck, R. (2007) A cost-based multi-unit resource auction for service-oriented grid computing. In 8th IEEE/ACM International Conference on Grid Computing (Grid 2007) Austin, Texas.
Skillicorn, D. B. (2002) Motivating computational grids. In 2nd IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2002), pp 401-406, IEEE Computer Society.
Xie, L., Smith, P., Banfield, M., Leopold, H., Sterbenz, J. P. G. and Hutchison, D. (2005) Towards resilient networks using programmable networking technologies. In IFIP IWAN 2005, Sophia-Antipolis, France.
Zweifel, P. and Eisen, R. (2002) Versicherungsökonomie. Springer, Heidelberg.