Cost reduction of wear-out monitoring by measurement point selection

Urban Ingelsson, Shih-Yen Chang and Erik Larsson
urban.ingelsson@liu.se, shich922@student.liu.se, erik.larsson@liu.se
Department of Computer and Information Science, Linköpings universitet, Sweden

Abstract - *Early failure rates have increased due to reduced feature dimensions and electromigration wear-out. Periodic delay measurements can be employed to estimate the state of wear-out. Including delay measurement sensors on-chip is costly. Therefore, a method is proposed to reduce the number of measurement points. The method identifies wear-out sensitive interconnects and selects a small number of measurement points to target the identified interconnects. The method is demonstrated on ISCAS85 benchmark ICs.*

I. Introduction

In the pursuit of faster, feature-richer and less power-consuming ICs, feature dimensions (transistor dimensions and interconnect width) are pushed further into the deep submicron region in each process generation. In this context, electromigration becomes relevant. Electromigration is a wear-out process which may cause ICs to fail in operation. The thinner an interconnect is, the sooner it may fail [1]. Wear-out monitoring aims at predicting impending failure due to wear-out.

Electromigration is known since the 1960ies. Recently, reductions in the cross-section area of interconnects have caused electromigration to become a significant issue [2]. The rate of interconnect wear-out was previously small in comparison with the interconnect cross-section area and the mean-time-to-failure ($MTTF$) was large compared with the expected period of use. With reduced cross-section area of interconnects, $MTTF$ has become short enough to impact the useful life-time of ICs.

Studies [3], [4], [5], [6], [7], [8] have focused on predicting the location of electromigration-induced damage to adapt the IC design accordingly. Other studies [9], [10], [11], [12], [13], [14] have proposed embedded wear-out monitors to predict the time of failure. The typical approach is to measure the delay through the combinatorial parts of the IC to observe propagation delay increase indicative of wear-out. Embedded wear-out monitors are expensive in terms of the silicon area. Whether the monitors are based on additional latches on the IC's flip-flops [11] or based on delay measurement circuits [13] the cost scales with the number of measurement points.

This paper suggests a low-cost wear-out monitoring solution with regard to electromigration. The method selects a low number of measurement points to target the interconnects that are sensitive to electromigration. Background on electromigration is presented in Section II. Previous approaches to wear-out monitoring are reviewed in Section III and the proposed wear-out monitoring solution is described in Section IV. A small set of measurement points is determined as described in Section V. Further, Section VI presents results based on ISCAS85 benchmark circuits and Section VII concludes the paper.

II. Electromigration Background

The wear-out mechanism electromigration describes gradual movement of metal atoms when influenced by a current [8]. High-momentum electrons transfer some of their momentum to the metal atoms. Electromigration only affects interconnects that are longer than a certain *Blech length* L, which depends on current density J (current I divided by interconnect cross-section area A), as explained in [15]. Most interconnects are longer than L and the following discussion applies to interconnects that are longer than the Blech length.

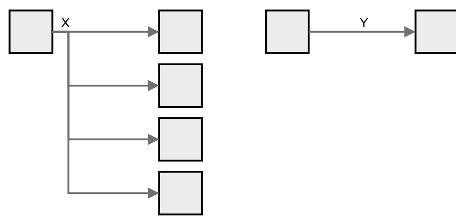
Electromigration can cause voids to form in the interconnect, often a via or a contact. The cross-section area is reduced by the void, and this leads to increased resistance and increased signal delay. With further wear-out, the void (and the resulting delay) will grow until the interconnect is broken and the void has developed into a full open defect. The effect is that ICs fail in operation after an unexpectedly short time.

According to [8], a narrow interconnect reduces the electromigration effect because of the grain structure of the material in the interconnect. An interconnect is not perfectly homogeneous but is made up of grains. Voids tend to form where the grains meet. An interconnect which is as narrow as the typical grain size has a reduced probability for voids to form. However, as described in [5], voids can form also within grains and narrow interconnect cannot prevent wear-out due to electromigration altogether.

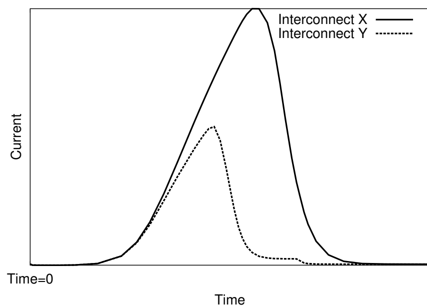
The rate of wear-out increases with the current density. In fact, the current density in interconnects during switching of logic states in CMOS ICs has been non-decreasing over the process generations. This is because supply voltage and interconnect cross-section area have historically scaled in similar rates. The result is an increase in wear-out rate and more importantly, the today's IC's interconnects have a smaller nominal cross-section area than in the previous decade. Therefore, $MTTF$ as described by Black's equation [1] (Equation 1) has decreased. Here, A is the original cross-section area, E_a is the activation energy specific to the material, J is the current density and T is the temperature. Further, k is Boltzmann's constant and n is a scaling factor.

$$MTTF = A \cdot (J^{-n}) \cdot e^{\frac{E_a}{k \cdot T}} \quad (1)$$

From Black's equation (Equation 1), it can be seen that the rate of wear-out increases with current density and temperature. The rate of wear-out also increases if current density is constant and the cross-section area is reduced. As the cross-section area A decreases, the resistance R of the wire increases as $R \propto A^{-1}$, and the current I decreases as $I \propto A$ (as can be seen by $I = V/R$), resulting in a constant current density $J = I/A$. Consequently, for a given interconnect and a constant voltage across the interconnect, the rate of wear-out on the interconnect is constant, even though the cross-section area is gradually reduced as a result of wear-



(a) Interconnect X and Interconnect Y



(b) Current waveforms

Fig. 1. Comparison of interconnects with high and low fanout

out. To mitigate electromigration, copper has been used to replace aluminum as material for metal interconnects, since copper is less susceptible to electromigration and has higher activation energy than aluminum [5]. However, the use of copper does not prevent electromigration altogether [5]. It should be noted that Black's equation (Equation 1) describes $MTTF$ under a constant flow of current. In fact, digital ICs in CMOS technology ideally has non-zero current only while switching between logic states. Therefore, the number of logic state switches that an interconnect has been subjected to influences its state of wear-out. This means that interconnects with high signal activity, i.e. where there are frequent switches of logic states, are particularly affected by electromigration.

The duration of current flow in each logic state switch has an impact on the wear-out accomplished. Figure 1 illustrates how an interconnect X with a fanout of four has a larger capacitive load than an interconnect Y with a fanout of one. The interconnect X will be subjected to a higher current for a longer period of time. In Figure 1(b), the currents through the driver of X and the driver of Y are shown for a logic switch initiated at Time=0. The reason for the longer-lasting current through X is that it takes longer time to charge or discharge a large capacitor than a small capacitor with constant supply voltage.

To summarize, the interconnects which are likely to be worst affected by electromigration are longer than the Blech length, have a small cross-section area, are more than a grain size wide, are subjected to much signal activity and have a large fanout. We call the electromigration-sensitive interconnects Wear-out Sensitive Nets (WSNs). This paper will describe how to identify WSNs and propose a method for finding the minimal number of measurement points for monitoring of the set of WSNs. The purpose is to reduce the cost of predicting failure due to electromigration.

III. Previous Work on Wear-Out Monitoring

In known approaches to wear-out monitoring [9], [10], [16], [11], [12], [13], [14], some consider electromigration [9], some consider NBTI (Negative Bias Temperature Instability) [11], [12], [14], some gate oxide breakdown [9], [13] or some other wear-out mechanism [16], [17]. Several wear-out mechanisms affect the propagation delay. Therefore, many wear-out monitoring approaches observe delay or slack. Four main approaches can be listed as follows. Firstly, prognostic cells [9] are designed to fail before the functional parts of the IC. When a prognostic cell fails, it is likely that the functional parts are also affected by wear-out, since they have operated under the same conditions. Prognostic cells can monitor wear-out without interrupting the IC function. However, it is not known to what extent the functional parts are affected by wear-out. Secondly, one type of wear-out monitors are based on regularly performed speed tests [10]. Reduction in system-wide slack indicates wear-out. The speed tests sample the IC state of wear-out while ascertaining that the monitoring circuitry is itself not worn-out prematurely [16]. During speed tests, the monitored IC must be in test mode. Therefore, the monitoring is scheduled for times in which the IC is idle. The speed test based approach cannot monitor short paths for which the nominal delay is short in comparison with the clock period. Thirdly, another type of wear-out monitors are based on augmented flip-flops for checking slack with regard to a guard-band [11], [12]. An approach was described in [11] with an adjustable guard-band, which tracks the smallest slack over all the monitored flip-flops. Some reduction in slack is tolerated and there is compensation by adjusting the supply voltage. This approach cannot monitor short paths. Fourthly, some monitors employ delay (or slack) measurements on the combinatorial paths [13]. In a sequential circuit, the combinatorial parts receive their inputs from flip-flop outputs or the primary inputs, and their outputs are flip-flop inputs or primary outputs. This approach requires on-chip delay measurement circuitry, which can be a significant silicon area overhead. A low-overhead delay measurement circuit was presented in [18]. A benefit of the approach with delay measurements is that short paths can be monitored explicitly. It should be noted that delay measurements are performed in test-mode and require predetermined input vectors. All of the four approaches for wear-out monitoring detailed above could be applied to target electromigration wear-out. There is a trade-off between the monitoring capability and overhead, with explicit delay measurements being the approach with the highest overhead but with ability to target all paths.

This paper considers an approach (Section IV) with explicit delay measurements to target interconnects that are particularly sensitive to electromigration wear-out. Section V describes how to keep down the cost of the delay measurement circuitry by identifying a small number of measurement points. For the wear-out mechanism NBTI, similar identification of sensitive circuitry has been done to aid wear-out monitoring, in [14], but this paper presents the first combination of identification of sensitive circuitry and wear-out monitoring for electromigration.

IV. Approach to Wear-Out Monitoring

To give a context to measurement point selection (Section V), we present a wear-out monitoring approach.

By regular, semi-periodic measurements of the time it takes for a signal to propagate through the circuit along a path that includes an interconnect, changes in that signal propagation delay can be attributed to wear-out. When the additional delay due to wear-out has increased beyond some value corresponding to severe wear-out, the wear-out monitor gives a warning to the system operator. The process from measurement to warning is described in Figure 2.

In step (1), a given pair of input vectors cause a signal transition at the input at the start of the clock period. The pair of input vectors set up a propagation path (thick arrows) so that the signal transition travels through a series of gates and interconnects to an output. The marked path includes two interconnects that are sensitive to electromigration wear-out. The identification of such interconnects is discussed in Section V. In step (2) in Figure 2, a measurement of the slack starts when the last transition occurs

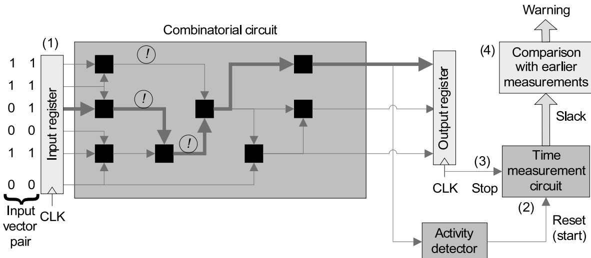
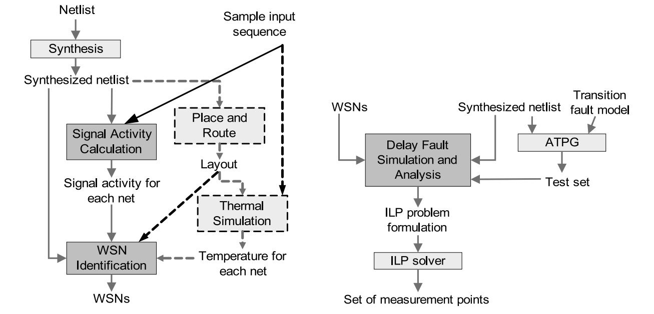


Fig. 2. Process from delay measurement to decision about giving warning



(a) Tool flow for WSN identification (b) Tool flow for measurement point selection

Fig. 3.

$$\text{minimize} \sum_{o \in Outputs} O_o$$
$$\forall_{w \in WSNs} \sum_{o \in Outputs} z_{w,o} \cdot O_o \geq 1$$
$$\forall_{o \in Outputs} 0 \leq O_o \leq 1$$

Fig. 4. ILP formulation

on the output. In step (3), the slack measurement ends when the clock period ends. A slack measurement at the output *cover* all interconnects that are on the path, and with other input vector pairs, the same output can cover interconnects in other paths that end at that output. Since the delay is added up, any increase in delay (reduction in slack) can be caused by wear-out on some of or all of the covered interconnects. The wear-out monitor gives a warning for a significant increase in delay disrespective of if the wear-out occurs on only one of the interconnects or on all of them. In the step marked (4) in Figure 2, the measured slack is compared against earlier measurements. If the comparison shows that the delay has increased significantly, a warning is given. Variations in temperature can temporarily affect the signal propagation delay through an IC [13]. To avoid giving false warnings the decision is based on the last few delay measurements. If no warning is given based on the measurement, the process continues by going back to step (1) to target another set of interconnects. When all considered interconnects have been covered or when a warning has been given, the delay monitoring process ends and is scheduled to restart at a later time.

The monitoring approach described above is based on explicit slack measurements, similar to the delay measurements in [13], adapted to the low overhead delay measurement circuit presented in [18]. The measurements are performed regularly, using pre-generated test vectors, while the monitored circuit is idle, at appropriate intervals to keep the monitors from wearing out before the monitored circuit, as in [16].

V. Method

The method for selecting a low number of delay measurement points to achieve a low-cost wear-out monitoring solution is divided in two phases, wear-out sensitive net (WSN) identification (Figure 3(a)) and measurement point selection (Figure 3(b)). A key idea is to place delay measurement points on the outputs of the combinatorial parts of the IC that is to be equipped with wear-out monitoring capability. The delay observed at these outputs includes the increased delay from interconnects that are worn-out due to electromigration. The method selects the appropriate set of outputs to equip with delay measurement points to target the wear-out sensitive nets.

The WSN identification phase is detailed in Figure 3(a). The steps *Signal Activity Calculation* and *WSN Identification* are specific to the presented method. The other steps, *Synthesis*, *Place and Route* and *Thermal Simulation*, correspond to off-the-shelf tools. For a design, described by a netlist, and a sample input sequence, the WSN identification step generates a set of WSNs and a synthesized netlist. These outputs from the WSN identification step will be used as inputs to the measurement point selection phase. As can be seen in Figure 3(a), the WSN identification phase proceeds with Synthesis to produce a synthesized netlist. The synthesized netlist is analyzed together with the sample input sequence to find the signal activity on each interconnect. This analysis can be implemented using a logic-level circuit simulator to identify the circuit state for each input vector. For two consecutive circuit states A and B, each interconnect i has two logic states A_i and B_i. If A_i is different from B_i, an activity counter for the interconnect is incremented. The signal activity is the activity count divided by the length of the sample input sequence v. To identify the WSNs, the synthesized netlist is analyzed together with the signal activities for the interconnects. If an interconnect has a fanout above a certain user-specified limit $FanoutLimit$ and a signal activity above another user-specified limit $SignalActivityLimit$, then the interconnect is regarded as a WSN.

The experiments presented in Section VI will show results based on a $SignalActivityLimit$ of 0.49 and a $FanoutLimit$ of two, as a proof-of-concept that it is possible to narrow down the number of targeted interconnects and for the targeted interconnects find a small number of measurement points. From Section II it can be subsumed that the identification of WSNs can be further refined by considering the Blech length and hot-spots. This requires place-and-route and thermal simulation. The procedure is marked with dashed arrows in Figure 3(a). With consideration of Blech length and hot-spots, one could limit the selection of WSNs to interconnects that are longer than the Blech length and reach a temperature above some limit $ThermalLimit$. The proper adjustment of the $FanoutLimit$, $SignalActivityLimit$ and $ThermalLimit$ is outside the scope of this paper. To demonstrate the method, the results in Section VI are based on a setup including only the interconnect signal activity and fanout for WSN identification.

The measurement point selection phase is detailed in Figure 3(b). Here, the step *Delay Fault Simulation and Analysis* is specific to the proposed method and *ATPG* and *ILP solver* are off-the-shelf tools. For the set of WSNs, the synthesized netlist and a transition fault test set generated by ATPG, an Integer Linear Programming (ILP) problem is formulated as shown in Figure 4. A solver, such as LP_Solve [19], is applied to arrive at a small set of delay measurement points for wear-out monitoring. Here, O_o is a variable that is 1 if Output o is selected as a measurement point and 0 otherwise. Furthermore, $z_{w,o}$ is 1 if a delay fault on WSN w can be observed on Output o, and 0 otherwise. To obtain the $z_{w,o}$ values, fault simulation is performed using the transition fault test vector pairs with and without an inserted delay fault on each WSN. The set of outputs for which the logic behavior depend on the existence of a delay fault are recorded in the $z_{w,o}$ values. This fault simulation is a key task of the tool flow for measurement point selection as seen in

TABLE I
FOUND WEAR-OUT SENSITIVE NETS IN ISCAS85 BENCHMARK DESIGNS

Circuit	Gates	Outputs	Inter-connects	Inter-connects with high activity	Inter-connects with large fanout	WSNs
C17	6	2	5	2	2	1
C432	160	7	173	42	61	2
C499	202	32	192	37	50	18
C880a	383	26	264	42	85	7
C1908	880	25	224	37	105	21
C2670	1269	140	337	112	113	34
C3540	1669	22	776	107	325	48
C7552	3513	108	844	381	503	186

Figure 3(b). The idea behind the ILP problem formulation is to find the outputs that can be used as measurement points for each WSN and then select the minimal set of outputs to target all WSNs. Note, the minimal set of measurement points is selected with respect to a given set of transition fault test pattern pairs and a given set of WSNs.

VI. RESULTS

To demonstrate how the proposed measurement point selection method can identify a small number of outputs for which delay measurements should be performed, experiments were performed on ISCAS85 benchmark circuits. The setup and the results of the experiments are described using Table I for detailing how wear-out sensitive nets (WSNs) are identified and Table II details how the WSNs are targeted by delay measurements on a number of outputs. In Table I, the four first columns show the name of the considered circuit, its number of gates, its number of outputs and its number of interconnects respectively. From the interconnects of each circuit, WSNs (right-most column) were identified based on the signal activity and the fanout. The signal activity was calculated based on a sequence of 1000 pseudo-random input vectors. Interconnects with signal activity higher than 0.49 transitions per clock cycle (fifth column) and a fanout of two or more (sixth column) were considered as WSNs. This means that the interconnects that are considered as WSNs change logic value about every other clock cycle and have more than the average capacitive load. As can be seen from Table I, less than half of the interconnects have such high signal activity and up to around 60% of the interconnects (in the case of c7552) can have such large fanout, and this depends on the circuit. It should be noted that the interconnects with high signal activity do not necessarily have large fanout, resulting in a small set of WSNs.

Targeting the WSNs that are described in Table I, the experiment resulted in a set of measured outputs as described in Table II. The first column shows the name of the considered circuit. To determine how the WSNs can be observed at the outputs, we used a set of transition fault test vector pairs, as described in Section V. The number of test vector pairs are shown in the second column. The proposed method selected a small set of outputs that can be used to observe all the WSNs, and the number of selected outputs is shown in the third column. The fourth column shows the number of measured outputs as a fraction of the total number of primary outputs (third column of Table I). The right-most column shows the computation time. The experiments were performed with in-house tools for calculating signal activity and for finding the outputs that can be used to target the WSNs. These tasks rely heavily on circuit simulation and fault simulation. With a state-of-the-art fault simulator to perform this task, computation time should be drastically reduced. Furthermore, computation time is not a very significant parameter, since measurement point selection is done during design time.

TABLE II
RESULTS WITH ISCAS85 BENCHMARK DESIGNS

Circuit	Test vector pairs	Measured outputs	Ditto as fraction of outputs	Computation time
C17	6	1	50%	6s
C432	45	1	14.3%	3min
C499	80	2	6.3%	7min
C880a	35	5	19.2%	5min
C1908	67	1	4%	8min
C2670	55	8	5.7%	24min
C3540	135	4	18.1%	1h 20min
C7552	72	15	13.9%	4h

The results in Table II show that compared to a wear-out monitoring approach that measures delay on all outputs, the proposed approach can reduce the number of measured outputs by 50% (in the case of C17) to 96% (in the case of C1908). Logic depth has an impact on the possible reductions in the number of measured outputs. For C17, the logic depth is 3 and for C1908, the logic depth can be 6 or above.

VII. CONCLUSION

This paper considered electromigration wear-out and presented a method for low-cost wear-out monitoring. The method selects a small number of measurement points for delay measurement, based on wear-out sensitive interconnects and outputs which can be employed to observe the wear-out state on the sensitive interconnects. By solving an ILP problem, a small set of outputs are selected. Experiments on ISCAS85 benchmarks showed 50% to 96% reduction in the number of measurement points compared to measuring on all outputs.

REFERENCES

[1] J. R. Black, "Mass transport of aluminum by momentum exchange with conducting electrons," in *Annual Reliability Physics Symposium*, Nov. 1967, pp. 148–159.
[2] L. Zhang, J. P. Zhou, J. Im, P. S. Ho, O. Aubel, C. Hennestal, and E. Zschech, "Effects of cap layer and grain structure on electromigration reliability of Cu/low-k interconnects for 45nm technology node," in *IRPS*, May 2010, pp. 581–585.
[3] S. Minehane, R. Duane, P. O'Sullivan, K. G. McCarthy, and A. Mathewson, "Design for reliability," *Microelectronics Reliability*, vol. 40, pp. 1285–1294, 2000.
[4] S. C. Choi and R. K. Iyer, "Wear-out simulation environment for VLSI designs," in *FTCS*, Jun. 1993, pp. 320–329.
[5] Y.-C. Joo and C. V. Thompson, "Electromigration-induced transgranular failure mechanisms in single-crystal aluminum interconnects," *J. of Appl. Phys.*, vol. 81, no. 9, pp. 6062–6072, 1997.
[6] Q. F. Duan and S. Y.-L., "On the prediction of electromigration voiding using stress-based modeling," *J. of Appl. Phys.*, vol. 87, no. 8, pp. 4039–4041, 2000.
[7] K. Sasagawa, M. Hasegawa, M. Saka, and H. Abe, "Prediction of electromigration failure in passivated polycrystalline line," *J. of Appl. Phys.*, vol. 91, no. 11, pp. 9005–9014, 2002.
[8] S. P. Hau-Riege and C. V. Thompson, "Electromigration saturation in a simple interconnect tree," *J. of Appl. Phys.*, vol. 88, no. 5, pp. 2382–2385, 2000.
[9] S. Mishra, M. Pecht, and D. L. Goodman, "In-situ Sensors for Product Reliability Monitoring," in *Design, test, integration and packing of MEMS/MOEMS*, May 2002, pp. 10–19.
[10] B. Zandian, W. Dweik, S. H. Kang, T. Punihaole, and M. Annavaram, "WearMon: Reliability Monitoring Using Adaptive Critical Path Testing," in *DSN*, 2010, pp. 151–160.
[11] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. C. Paul, W. Wang, B. Yang, Y. Cao, and S. Mitra, "Optimized Circuit Failure Prediction for Aging: Practicality and Promise," in *ITC*, Oct. 2008, pp. 1–10.
[12] J. C. Vazquez, V. Champac, I. C. Teixeira, M. B. Santos, and J. P. Teixeira, "Programmable Aging Sensor for Automotive Safety-Critical Applications," in *DATE*, Mar. 2010, pp. 618–621.
[13] J. Blome, S. Feng, S. Gupta, and S. Mahike, "Self-calibrating Online Wearout Detection," in *International Symposium on Microarchitecture*, 2007, pp. 109–120.
[14] A. H. Baba and S. Mitra, "Testing for transistor aging," in *VTS*, 2009, pp. 215–220.
[15] S. P. Murarka, I. A. Blech, and H. J. Levinstein, "Thin-Film Interaction in Al and Pt," *J. of Appl. Phys.*, vol. 47, no. 12, pp. 5175–5181, 1976.
[16] Y. Li, S. Makar, and S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test Usign Stored Test Patterns," in *DATE*, 2008, pp. 885–890.
[17] Y. Sato, S. Kajihara, Y. Miura, T. Yoneda, S. Ohtake, M. Inoue, and H. Fujiware, "A Circuit Failure Prediction Mechanism (DART) for High Field Reliability," in *ASIC*, Oct. 2009, pp. 581–584.
[18] S. Pei, H. Li, and X. Li, "A Low Overhead On-chip Path Delay Measurement Circuit," in *ATS*, 2009, pp. 145–150.
[19] http://lpsolve.sourceforge.net/5.5/, Mar. 2011.