Resolving the Trade-Offs in Designing QoS Communication Services for Control Applications on CAN

Jörg Kaiser *
Edgar Nett **
*University of Ulm, Germany[1]
kaiser@informatik.uni-ulm.de
** Otto-von-Guericke-University Magdeburg, Germany
nett@ivs.cs.uni-magdeburg.de

Abstract

Systems designed for control applications typically have to consider quality and safety requirements. Quality issues allow a smooth and convenient control while safety constraints try to prevent any dangerous situation. While quality issues can be treated on a statistical basis, safety constraints have to be guaranteed in each individual case. Because the requirements often are handled statically, quality and safety constraints can not be separated properly. This results in a unnecessarily high overall resource demand. The paper discusses mechanisms which allow to treat quality and safety properties of a communication system separately. The basic approach is to identify the strictly safety related properties of the application. These properties are guaranteed by static reservations. The remaining resources are shared on the basis of a deadline driven mechanism. By monitoring the system, we also can reclaim statically reserved resources if the respective message transmission is not needed from the safety point of view . The mechanisms are embedded in the concept of event channels which allow to reflect the properties of the underlying communication system.

1. Introduction

Technological advances enable distributed control systems composed from cooperating smart sensor and actor components. The advantage is that computing and signal processing can be distributed to the components. However, the communication system constitutes the central shared resource for information dissemination. One of the most important issue for such a communication system is the trade-off between safety and quality requirements. While quality issues can be treated on a statistical basis, safety constraints have to be guaranteed in each individual case. Therefore, safety considerations may impose considerable constraints on the throughput by requiring time redundancy like the inclusion of multiple message retransmissions to tolerate transmission faults. On the other side, it is desirable to have as much bandwidth available as possible to improve the quality of control. In many control applications, the sampling rates to perceive the environment adequately are adjusted for fine grained control rather than merely meeting safety constraints [1] resulting in high message rates. Safety constraints would allow for less frequent sampling periods. Therefore, if critical and less critical messages could be distinguished, bandwidth resulting from the more expensive transmission of critical messages could be saved. The problem is that safety properties have to be guaranteed which requires some form of off-line analysis and planning. At least, the planning cannot be in the time-critical path of an application. This results either in a static schedule based on priorities (like deadline/rate monotonic DMS/RMS) or time slots (like TDMA), or, alternatively, stringent assumptions about inter-arrival times to avoid transient overload situations for flexible deadline based planning [2]. In a purely time-triggered or RMS-based scheme the distinction between different classes of messages cannot be realized.

This paper suggests a solution for this throughput-safety trade-off which has been implemented for the popular CAN-Bus. Instead of preventing temporal

[1] This work has partly been supported by EU under research contract IST-2000-26031 (CORTEX)

faults for every message transmission by worst-case assumptions, we propose a scheme inspired by fault-tolerance. We strive for tolerating a number of temporal faults while preserving safety properties. The main idea is to identify application-inherent redundancy with respect to safety, which is available in most control systems because of quality reasons. Our scheme thus puts the fault-tolerance approach to a new perspective of flexible scheduling in real-time systems. It exploits the redundancy in the normal case for quality improvement while securing the minimum functionality

Our work differs from research in flexible scheduling [2] in that it explicitly considers faults and transient overload in the communication system. There are a number of related proposals which tackle the principle problem of the throughput/safety trade-off and deal with predictability not for each individual message but allow to tolerate a number of deadline misses. What is important in our work, is that a bound on missed deadlines can be established and enforced by the system. This is similar to the (m,k)-fim real-time concept of Hamdaoui and Ramanathan [3] used to schedule a communication system. However Hamdaoui and Ramanathan do not provide any mechanism for guarantees. Bernat and Burns [4] presented the conceptual framework of weakly-hard real time systems. Our approach has similarities to their scheduling of dual priorities. Any interference of critical communication instants is avoided by a reservation scheme. These messages are receive the highest priority when they become ready and thus their dissemination is guaranteed. Recently, this approach has been studied in a CAN environment [5]. However, we firstly use a TDMA reservation scheme instead of priority scheduling, secondly we use deadlines to schedule the non critical traffic as a way to assign individual temporal properties also to aperiodic messages without a central aperiodic server [6] or a centralized scheduler [7]. Thirdly, we provide an efficient way of dynamically reclaim unused bandwidth of scheduled TDMA slots. Additionally, we embed our scheme in the concept of event channels which allows to describe the temporal requirements of communication on a higher abstraction level than just messages.

2. Exploiting Application Specific Redundancy

Application-inherent redundancies can find their expression in several ways. Many control applications exhibit timing redundancy in how often certain modules have to be executed within a given time. The frequency at which controllers are executed is usually chosen to be significantly higher than would be necessary for a safe operation of the system. This is because two considerations guide the selection of the frequency: It must be sufficiently high so that the controller (i) can react to changes in the controlled system before it is damaged (a safety constraint) and (ii) exhibits a smooth reaction to the changes in the controlled system (a quality goal). While (ii) is less critical regarding the safety of the systems, it implies the more stringent timing requirements.

An example of such a kind of controller is used to control the probe of an Atomic Force Microscope. Atomic Force Microscopy allows scanning the surface of a specimen with a very high resolution. To achieve a quasi-continuous control of the probe-to-specimen distance and a high quality of the scan, the controller executes at a frequency of 100kHz. Applying Nyquist's theorem or Shannon's law to the maximum frequency to be expected in the system yields a frequency of 23 kHz, but choosing a higher frequency significantly improves the quality of the results. In fact, even a reduction from 100 kHz to 50 kHz leads to perceptible worse results. The worst-case response time of the controller to avoid damage of the probe can be computed taking into account the speed of the specimen, the topography of the specimen, and the range of the sensors. It turns out that a worst-case response time of 13,5ms is sufficient to ensure the safety of the system. Thus, the chosen frequency of 100kHz exceeds the minimum frequency required by a factor of 1350. The over-sampling is a typical technique in control systems and therefore can also be observed in many other application.

The second example addresses the problem of more dynamic systems like a team robot application. In this application, it is desirable to exploit the diverse and complementary sensors to improve the environment perception and to extend the range of sensing. Cooperation is performed on the basis of local sensors and communicated events which may carry remote sensor events and the necessary control information to coordinate actions. It is obvious that the cooperation aspect introduces a predictability problem. In our example, a robot equipped with a with line tracking camera guides a "blind" vehicle only equipped with distance sensors to detect obstacles and comparing the sensor readings with those of the guide robot it eventually can detect the guides presence. Because the blind vehicle can exploit the remote sensor information of the guide, it can follow the guide reliably even at high speeds. A couple of problems arise because of the dynamic coupling and wireless information

dissemination. Due to the dynamic nature of the interaction, any a priori statically planned dissemination schedule is impossible. Secondly, wireless channels have to cope with a high number of transmission faults. Worst-case assumptions severely would constrain the possible throughput. In the example, the safety properties for the robots are not crash and not to loose each other. The application specific redundancy comes from the fact that for a short time and lower speed these properties can be handled by the local distance sensors only. Therefore again, there is potential to trade quality and safety parameters if flexible mechanisms are available.

3. Handling Temporal Specifications by Flexible Mechanisms

There are three aspects related to exploit the application inherent redundancy: Firstly, we need some interface to the application where we can specify the temporal requirements easily. Secondly, we have to provide a scheduling strategy which allows to handle the trade-off between reliability and predictability as well as the safety/throughput problem. Finally, there must be mechanisms on the network level to enforce the specified properties.

We adopted an event-based communication model to describe all interactions. The scheme has been implemented in the COSMIC middleware (COoperating SMart devICes) [8] for the CAN-Bus widely used in automotive industry. Events allow a high level specification of temporal properties related to an individual occurrence e.g. the temporal validity of a event. Events are disseminated in a publisher subscriber style through event channels. The notion of an event channel allows to model the quality of the communication system. We provide event channels of different synchrony classes. Hard real-time event channels always deliver an event at the defined deadline. Any interference between hard real-time channels is omitted by using a TDMA scheme and static analysis. Because hard real-time channels guarantee delivery under certain classes of transient faults, retransmissions and temporal unavailability of the communication medium has to be considered. This extends the length of a reserved slot to a multiple of a single CAN message and therefore, depending on the fault model, the number of possible TDMA slots drops down to 350 slots/sec compared to a maximum throughput of about 6500 maximum length messages on a CAN-Bus with a transfer rate of 1Mbit/sec. Further, it should be noted that, in general, sporadic safety critical events must be mapped to a periodic scheme to guarantee predictable dissemination as it is common in the TDMA approach. In a conventional TDMA scheme these slots cannot be reused by less critical traffic if no critical message has to be sent. In COSMIC the reservation of slots is enforced by the CAN priority scheme. This allows lower priority messages to be sent automatically if no critical message is ready. Moreover, it is possible to determine dynamically if a message has been received successfully by all subscribers. Bandwidth allocated for redundant retransmissions which is not needed can now be used by the less critical traffic. This allows very conservative worst-case assumptions because the penalty comes in effect only if the worst case really happens.

The problem which remains is that the number of slots is restricted. Here the combination with the TAFT (Time-Aware Fault-Tolerant) [9] scheduling approach allows to add another degree of scalability. A total number of 350 usable slots on a CAN-Bus may result in a serious problem if multiple periodic hard real-time event channels have to be reserved. Therefore, exploiting the application inherent redundancy can ease this problem substantially and in many cases make it at all possible to guarantee the safety requirements maintaining at the same time the desired quality properties. TAFT provides predictability guarantees in a system in which only k messages from a total of m messages have to be sent successfully. Because of this, the number of critical hard real-time messages is reduced and also the number of time slots which have to be reserved. This eases the task to find an appropriate allocation. Secondly, because now some of the previous hard real-time events can be moved to a soft real-time class, they compete with other sporadic soft real-time events on the basis of deadlines which increases the probability for the overall message set to minimize lateness. Additionally, if a periodic event has been disseminated k times successfully trough a soft real time channel, the allocated hard real-time slot can remain unused and thus, other events can use the available bandwidth. In a purely time-triggered system, this would not be possible. The properties of TAFT are presented in detail in [10].

A first experiment to evaluate the behaviour of the protocol uses a system of 3 CAN nodes running the COSMIC under RTLinux. The bit rate of the CAN-Bus was adjusted to 125 kbits/sec. All messages have a payload of 8 bytes thus resulting in a message transfer time of 154 bit times =1232 μsec. The maximum (100%) load therefore is 811 messages/sec. This results in a slight overload if the slots of the hard real-time event channels would be used completely.

We arranged the HRT message slots with an omission degree of 3 and a period of 8.6 ms. This results in a time slot length of 815 bit-times (6.52ms). The HRT slots cover 61.5% of the entire bandwidth. Note that in purely time-triggered schemes this only would leave 38.5% for the remaining 2 messages (period 4.3 ms, 57,2% total band-width share). Because our scheme can exploit the bandwidth that is not used by HRT traffic, we have a network utilization of only 14.3% by HRT messages in the fault-free case. However, because of the critical phasing of the messages and tight deadlines for the soft real-time messages, the number of deadline misses still are high (Table 1a).

Table 1: Missed deadlines

	HRT	SRT1	SRT2
a.) (1-of-1)	0%	40%	10%
b.) (1-of-4)	0%	19%	8%

If we apply the TAFT approach and only reserve a hard real-time slot for 1 out of 4 messages while the other compete on a deadline basis, the overall deadline misses drop substantially (Table 1b). The hard real-time messages, because of the longer period, still do not miss any deadline. The experiments give an indication about the savings but larger experiments are planned to provide statistically relevant data.

4. Conclusion

The paper brings together previous work on TAFT (Time Aware Fault-Tolernat Scheduling) and COSMIC ((middleware for) Co-Operating SMart devICes). TAFT that was designed for flexible CPU scheduling contributed the basic ideas of handling the average case efficiently without sacrificing predictability. COSMIC allows coexistence of multiple real-time communication classes and thus provides the flexibility to put TAFT mechanisms to work. We proposed a k-out-of-n mechanism in which the safety constraints are met, if at least k message of every n messages are transferred successfully. A first evaluation shows that the overall throughput is increased and the percentage of missed deadlines for soft real-time messages is decreased in highly loaded networks where transient overload situations may occur frequently.

Future work will also include the application of the scheme to wireless communication nets. Here, even k of m guarantees may not always be possible. Our team robot example shows that in these cases, we may need application specific dynamic QoS adaptation [11] to meet the safety constraints. We will further investigate how the concepts presented in this paper will be exploited in such a scenario.

5. References

[1] J. C. Cunha, R. Maia, M. Z. Rela, J.G. Silva: "A Study of Failure Models in Feedback Control Systems", The International Conference on Dependable Systems and Networks (DSN), Göteborg, Sweden, 1-4 July 2001

[2] P. Marti, J.M. Fuertes, G. Fohler, K. Ramamritham "Improving Quality-of-Control using Flexible Timing Constraints: Metric and Scheduling Issues", 23rd IEEE Real-time Systems Symposium, Austin, TX,USA, Dec. 2002

[3] M. Hamdaoui and P. Ramanathan. A dynamic priority assignment technique for streams with (m, k)-firm deadlines. IEEE Transactions on Computers, 44(4), Dec.1995.

[4.] G. Bernat and A. Burns. Weakly_Hard real-time systems, IEEE Transactions on Computers, 50(4),pp.308-321,2001.

[5]I. Broster Flexibility in dependable Real-time Computing, PhD-Thesis Dept. of Computer Scienc, University of York, August 2003

[6] Nolte T., Nolin M., Hansson, H. Server-Based Real-Time Communication on CAN. *Proceedings of 11th IFAC Symposium on Information Control Problems in Manufacturing,* Salvador, Brazil, 2004

[7] L. Almeida, "A Word for Operational Flexibility in Distributed Safety-Critical Systems", in *Proc.8th IEEE Workshop on Object-Oriented Distrributed Systems*, Guadalajara, Mexico, January 2003

[8] Jörg Kaiser, Carlos Mitidieri, Cristiano Bruna, Carlos Eduardo Pereira: "COSMIC: A middleware for event-based interaction on CAN", ETFA, Emerging Technologies and Factory Automation, Lissabon, Portugal, 16.0-19.9. 2003

[9] Nett, E., Gergeleit, M., Mock, M., 2001: „Enhancing O-O Middleware to become Time-Aware", Special Issue on Real-Time Middleware in Real-Time Systems, 20 (2): 211-228, March, Kluwer Academic Publishers. ISSN- 0922-6443

[10] Nett, E. and S. Schemmer: „Reliable Real-Time Communication in Cooperative Mobile Applications", IEEE Transactions on Computers, 52(2), 2003, pp. 166-180.

[11] P. Verissimo, A. Casimiro, "Using the Timely Computing Base for Dependable QoS Adaptation." In Proceedings of the 20th IEEE Symposium onReliable Distributed Systems, New Orleans, USA, October 2001