Resolving the trade-offs in designing QoS communication services for control applications on CAN

Extended Abstract

Jörg Kaiser
University of Ulm
Ulm, Germany
kaiser@informatik.uni-ulm.de

Edgar Nett
Otto-von-Guericke University
Magdeburg, Germany
nett@ivs.cs.uni-magdeburg.de

1. Introduction

The communication system of a distributed control systems operating in an open dynamic environment has to cope with multiple conflicting needs. It has to guarantee the message dissemination under the strict safety constraints that are imposed by the interaction with the physical environment. These guarantees result in substantial resource requirements because they are based on worst-case assumptions and on the prevention of contention and overload situations. To guarantee proper dissemination of safety critical messages, any unpredictability because of using a shared communication medium has to be resolved. The most obvious conflicts can occur during arbitration among multiple nodes ready to send a message. Contention-based scheduling schemes give priority to sending messages with shorter deadlines and achieve a good usage of network bandwidth, but suffer from unpredictable delays in sending messages due to collisions. Contention free scheduling techniques, such as the TDMA (time division multiple-access) approach, avoid any occurrence of collisions. However, the pure TDMA results in a very poor utilization of network bandwidth. Time slots allocated for sending messages must comprise the longest time needed to send a message successfully and hence, consider retransmission of messages up to the anticipated omission degree. Thus, the TDMA approach has an inherent reliability/predictability trade-off.

As a second crucial problem, we have to consider the trade-off between throughput and safety requirements. In many control applications, the sampling rates are adjusted striving for fine grained control rather than merely meeting safety constraints. Therefore, safety constraints would allow for less frequent sampling periods. If critical and less critical messages could be distinguished, bandwidth could be saved. In a purely time-triggered scheme such a distinction cannot be realized. Because a time-triggered approach needs to reserve a slot for every periodic message and these slots are strictly assigned to a certain node.

The contribution of the paper is to suggest a solution for this throughput-safety trade-off particularly for the popular CAN-Bus, which predominantly is used in automotive applications. Instead of preventing temporal faults 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 and exploit application-inherent redundancy, 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.

There are a number of proposals which tackle the principle problem to relax the stringent requirements of hard real-time systems concerning the throughput/safety trade-off and to deal with predictability not for each individual deadline but allow to tolerate 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 [HaR95] which they use to schedule a communication system. However they do not provide any mechanism for guarantees. Bernat and Burns [BeB01] presented the conceptual framework of weakly-hard real time systems. Our approach has similarities to their scheduling of dual priorities. Any interference of critical tasks is avoided by a reservation

scheme. These tasks are mapped to the highest priority when they become ready and thus their dissemination is guaranteed. However, we firstly use a flexible TDMA reservation scheme instead of priority scheduling and secondly also provide an efficient way of dynamically reclaim unused bandwidth. Additionally, we provide a high level abstraction to describe the temporal requirements of communication.

2 Exploiting application specific redundancy

Application-inherent redundancies can find its 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 the example which we implemented, a robot equipped with a with line tracking camera guides a "blind" vehicle without such sensors. 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 interaction, any a priori statically planned dissemination schedule is impossible. Secondly, wireless channels have to cope with a substantially higher 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 these properties can be handled by the local sensors only. Therefore, there is potential again 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 a mechanisms on the network level to enforce the specified properties

We adopted an event-based communication model to describe all interactions. Events are disseminated in a publisher subscriber style. The term "event-based" in this context does not refer to any specific model of synchrony. Events may be spontaneously generated and immediately disseminated or periodically triggered by a clock. Events allow a high level specification of temporal properties related to an individual occurrence e.g. the temporal validity of a event. Additionally, we introduce the notion of an event channel to model the quality of the communication system. This allows the use of different synchrony classes for channels over which the events are disseminated. The scheme has been implemented in the COSMIC middleware (COoperating SMart devICes) [KMB03] for the CAN-Bus widely used in automotive industry. COSMIC allows the reuse of the bandwidth of real-time event channels. Hard real-time message transmission must be certain under a number of assumed omissions. Any interference between hard real-time channels is omitted by using a TDMA scheme and static analysis. Therefore, on a CAN-Bus, 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. This may be even worse in a wireless network. We further assume that, in general, even sporadic safety critical events must be mapped to a periodic scheme, to guarantee predictable dissemination by a statically reserved message slot. In a conventional TDMA scheme these slots are wasted completely if no message has to be sent. In COSMIC the reservation of slots is enforced by a priority scheme allowing other lower priority messages to be sent automatically if no critical message is sent in the reserved slot,. Moreover, it is possible to determine dynamically if a message has been received successfully by all subscribers and bandwidth which is not needed because there were no omissions, can 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) [NGM01] 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 [NeS03].

7. Conclusion

In our work we explore the trade-offs when designing communication services for control applications. Distributed safety critical applications require hard guarantees from the communication system. These guarantees are costly and occupy a substantial bandwidth of the scarce communication resources because worst case assumptions have to be made. However, in most control applications there is a large gap between what is needed for safety reasons and what is desirable because of quality reasons. We discussed this typical trade–off using the example of a scanning microscope and a team robot scenario. From the safety point of view we need predictability, from a quality point of view throughput would be the goal. Most of the current protocols for control systems focus on the safety point of view and are designed to avoid any source of unpredictability. They exploit time redundancy for forward error recovery and avoid contention of the shared communication

medium by a pre-planned approach. This conservative mechanisms incur severe bandwidth reductions and throughput degradations.

The basic idea of our approach is to apply mechanisms of fault-tolerance for a more cost-effective solution. First, we try to distinguish between the requirements imposed by safety constraints and the requirements that are defined by the desired quality of a control task. While resources for safety critical communication have to be guaranteed, quality issues can be handled on a statistical basis. The second source of usable bandwidth comes from resources that are not used because the worst case did not happen. This may be well over 90% of all cases. The reuse requires the dynamic detection of these resources. The dynamic priority-based arbitration scheme of the CAN-Bus in combination of the COSMIC mechanism can do this with a low overhead.

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 [VeC01] to meet the safety constraints. We will further investigate how the concepts presented in this paper will be exploited in such a scenario.

References

[KMB03] 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

[NeS03] Nett, E. and S. Schemmer: „*Reliable Real-Time Communication in Cooperative Mobile Applications",* IEEE Transactions on Computers, 52(2), 2003, pp. 166-180.

[NGM01] 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

[BeB01.] G. Bernat and A. Burns. *Weakly_Hard real-time systems*, IEEE Transactions on Computers, 50(4),pp.308-321,2001.

[VC01] 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

[HaR95] 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.

.