Software fault detection and recovery in critical real-time systems: An approach based on loose coupling

Pekka Alho*, Jouni Mattila

Department of Intelligent Hydraulics and Automation, Tampere University of Technology, Finland

HIGHLIGHTS

- We analyze fault tolerance in mission-critical real-time systems.
- Decoupled architectural model can be used to implement fault tolerance.
- Prototype implementation for remote handling control system and service manager.
- Recovery from transient faults by restarting services.

ARTICLE INFO

ABSTRACT

Remote handling (RH) systems are used to inspect, make changes to, and maintain components in the ITER machine and as such are an example of mission-critical system. Failure in a critical system may cause damage, significant financial losses and loss of experiment runtime, making dependability one of their most important properties. However, even if the software for RH control systems has been developed using best practices, the system might still fail due to undetected faults (bugs), hardware failures, etc. Critical systems therefore need capability to tolerate faults and resume operation after their occurrence. However, design of effective fault detection and recovery mechanisms poses a challenge due to timeliness requirements, growth in scale, and complex interactions. In this paper we evaluate effectiveness of service-oriented architectural approach to fault tolerance in mission-critical real-time systems. We use a prototype implementation for service management with an experimental RH control system and industrial manipulator. The fault tolerance is based on using the high level of decoupling between services to recover from transient faults by service restarts. In case the recovery process is not successful, the system can still be used if the fault was not in a critical software module.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Remote handling (RH) systems are used to inspect, make changes to, and maintain components in the ITER machine. Failure in a mission-critical system like RH may cause damage and, perhaps even more significantly, loss of experiment runtime, therefore making dependability one of its most important properties. However, even if the software for the RH system has been developed using valid development processes, the system might still fail due to undetected faults, hardware failures, etc. Critical systems therefore need to be able to resume operation after faults have occurred, but design of effective fault detection and recovery mechanisms poses a challenge. This is due to timeliness requirements combined with growth in scale and complex dynamic interactions in RH systems and embedded systems in general.

Several programming languages and frameworks, e.g. Erlang or OSGi for Java, support use of decoupled architectural models that can be used to implement fault tolerance solutions and dynamic loading of software modules, but these approaches are typically used in non-critical applications that do not have requirements for deterministic response times. In this paper we evaluate effectiveness of the decoupled architectural approach in mission-critical real-time systems using an experimental RH control system for an industrial manipulator. The control system is based on a real-time service oriented architecture (RTSOA) that we have introduced and evaluated in [1]. Services (i.e. the applications that participate in the control of the manipulator) are managed by a prototype

service manager that is used to detect faults and initiate recovery processes.

The RH control system consists of several heterogeneous subsystems, including equipment controller (EC), virtual reality (VR) and operations management system (OMS), specified in the ITER RH control system handbook [2]. This kind of cooperation of several networked computational units is typical for the field of cyber-physical systems (CPS), featuring a tight coordination between computational and physical elements of the system. CPS research aims to improve interoperability and openness between networked controllers to produce more intelligent applications.

2. Background

Fault tolerance means avoiding service failures in the presence of faults, and consists of error detection and recovery [3]. However, recovery can introduce unpredictable delays that might be in conflict with the predictability requirements of real-time systems. A typical approach for real-time fault tolerance is to use two or more diverse versions of software. Such an approach is suitable, e.g. in aviation, where the scope of critical systems is limited, and the cost of creating multi-version software is distributed over a large number of aircraft [4]. However, for large and complex one-off software systems, such as RH, use of multi-version techniques is difficult to justify.

Key challenges for fault tolerance in RH systems include recovery of state data, reliable detection of faults, fault recovery that supports real-time requirements, and ensuring reliability of end-to-end service chains. Safety of the system relies heavily on reasoning about consequences of faults, which is an important open research area due to the complex and stochastic nature characteristic for CPS.

Previous research on real-time fault tolerance has focused largely on redundancy-based solutions and reconfiguration. Gonzales et al. use adaptive management of redundancy to assure reliability of critical modules by allocating as much redundancy to less critical modules as could be afforded, thus gracefully reducing their resource requirements [5]. Assured reconfiguration in case of failures is used in [6]. This allows the primary function to fail and then reconfigure to some simpler function – reconfiguration of the system is a critical part, and it is formally verified. Simplex architecture by Sha et al. uses high assurance and high performance control subsystems [7]. The high assurance subsystem is used to keep the system within the safety envelope. ORTEGA architecture improves the Simplex architecture by adding on-demand detection and recovery of faulty tasks [8]. An anomaly based approach for detecting and identifying software and hardware faults in pervasive computing systems is proposed in [9]. The methodology uses an array of features to capture spatial and temporal variability to be used by an anomaly analysis engine.

Our work differs from these approaches by focusing on transient faults in a real-time system by using highly modular approach instead of redundancy. Similar solution based on modularity and fault isolation has been successfully used, e.g. in the non-real-time MINIX operating system (OS) for driver management [10].

3. Fault detection and recovery in real-time systems

3.1. System definition

Our hypothesis is that mission critical real-time systems can use service management to recover from transient faults (discussed in Section 5) in a loosely coupled software architecture. In this context, we define a loosely coupled real-time system as follows:
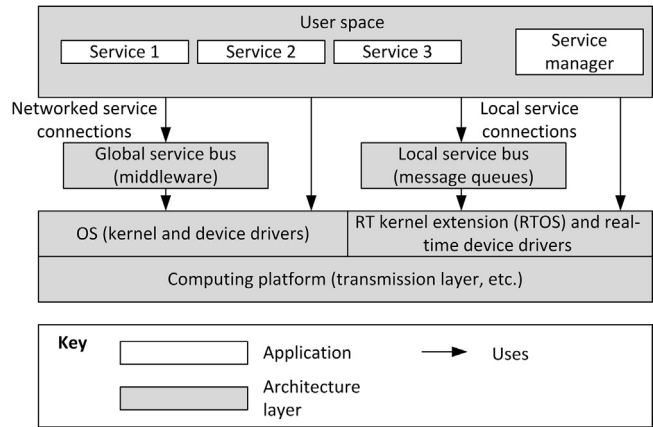

Fig. 1. Logical system architecture.

- The system is made up of a set of periodic processes, i.e. services.
- Services are loosely coupled, having no direct interdependencies or references to each other.
- Services can be distributed over network or located on a single computer.
- Services communicate with a communication buses that facilitate monitoring of communication deadlines.

Advent of modern, powerful processors to RH systems provides a chance to mitigate the delays caused by the recovery process if the fault is detected before deadlines. In an optimal case, a fault can be detected and recovered before it causes service failures. If fault recovery causes exceeding of a deadline, other services can detect this and react accordingly by moving the system to a safe state while simultaneously isolating the fault. Since this recovery strategy does not rely on redundant versions, there is no need to maintain consistency between replicas, which is a major challenge for redundant systems [11]. We also leave formal methods out of the scope of our solution because of architectural limitations and costs associated with these methods are likely to be prohibitive for ITER.

3.2. Architecture

The system architecture in our implementation is based on RTSOA using data-centric middleware and an open source real-time operating system (RTOS). It provides decoupled connections for the services via local and global service buses (LSB and GSB) and includes a service manager to monitor and manage services (Fig. 1) [1].

The LSB is based on real-time queues, a communication mode provided by the RTOS. A message queue can be created by one service and used by multiple services that send and/or receive messages to the queue. GSB is a wrapper that uses Data Distribution Service for Real-Time Systems (DDS) middleware for networked connections; DDS is a standard for decentralized and data-centric middleware based on the publish/subscribe model and aimed at mission-critical and embedded systems. The standard is maintained by the Object Management Group (http://portals.omg.org/dds/).

3.3. Service manager and service configuration

Service manager is a local component used to start services and detect faults. Service manager spawns the services as child processes, according to a configuration file – more advanced configuration methods could be supported, e.g. GUIs, web interface

or GSB-based, but have not been included in the prototype implementation at this phase. Services update their status to the service manager through either LSB or GSB. Possible states are *running, stopped, (re)starting*, and *error*. The local communication bus is also used by the service manager to command services to switch states (*start*, *stop* and *restart*). More extensive set of states and commands would give a more fine-grained control over services, but also add additional complexity to service implementation and service management logic, potentially introducing new faults.

Fault detection utilizes several methods: service status updates, monitoring user-set resource limits for CPU and memory usage, heartbeat monitoring, and OS features (parent process can check the state of its child processes). Service manager can be used to dynamically update services by terminating them and replacing with new version, enabled by the decoupled bus-based architecture.

Unresponsive behavior or unexpected increase in resource usage for a service can indicate a fault in the service, potentially endangering real-time performance of other services or causing dangerous movements – this is a typical challenge for design of CPS, needing interdisciplinary co-operation of designers and researches from industry and academia. After a fault has been detected, the service manager either terminates or restarts the faulty service, according to the configuration. These actions will either recover the fault without cascading service failures or put the system to a safe state, according to "let it fail" approach. This method is mainly suited to handling of transient faults. If the fault recovery fails, system simply stays in the safe state.

The fault recovery strategy enables operator to continue operation or recover equipment by avoiding the conditions that trigger the fault, postponing software update to later time and thus retaining higher system availability. Alternatively, operations can be suspended until the fault has been removed.

Effectively the described method of fault tolerance is a form of fault masking. The system therefore needs an error manager component to track number of errors so that they do not go unnoticed. This functionality is logical to implement in the service manager. Error data provided by the service manager can be used to give, e.g. graphical warnings to the user about encountered faults or send notifications via email or another message channel. Finally, if restarting a service does not solve the fault, service manager can utilize escalation and move the system to limp-home or recovery mode.

Configuration of services to be started and managed by the service manager may include the following parameters – new parameters can be easily added on per-need basis (parameters in *italics* are not implemented in the prototype):

- Start command, including necessary command line parameters.
- CPU and memory usage limits in percentage (*or bytes*).
- Heartbeat timeout.
- Actions to be taken on failure: restart, terminate, *execute {program}, alert {email address}*.
- Limit for the number of restarts.
- List of dependencies if they must be also restarted.

3.4. Service design

In the case of a failure, service stops execution or is terminated by the service manager and its state data will be lost. State data can be divided to temporary, static and dynamic: temporary data is related to current computations and is not relevant after a failure, static data is typically configuration data that can be reread, and dynamic data contains results of calculations, user input, etc. Some of the dynamic state data can also be recalculated or reread (e.g. sensor measurements), but commonly it needs to be protected. This means that any state data that needs to be recovered after restart must be stored in a stable storage. However, since it is possible that the saved state data has been corrupted, sanity-checking (typically checksums or valid range for data) is needed. Another issue regards how much of the state should be recovered. For example, an interrupted trajectory of a remotely operated manipulator should not be continued because of potential risks. The software may also hit the exactly same bug again if the full state from the time of failure is recovered.

Another major consideration for service design is that services should be fail-stop or fail-silent so that a failure in one application does not cause unwanted behavior in others. Compared to heartbeat timeouts and increased resource use, detection of erroneous outputs is more difficult to implement and has to be done using more "traditional" methods including contract programming, asserts, exceptions, etc., according to the needs of the specific application. If the service detects an internal error (e.g. an exception), and is still in otherwise sane state, it can report the error to the service manager by publishing status change to *error*. In our implementation, services publish the status in the heartbeat signal periodically sent to the service manager.

Service dependencies may be unavailable at times due to faults or even normal use scenarios. Service developer therefore needs to implement monitoring for communication deadlines and decide how to recover services from faults. Recovery should be taken into account already in the specification [12]. This is a direct consequence of the dynamic nature of decoupled architectures and forces the developer to take into account situations where the dependencies of a service are offline or unavailable. Although the dynamic service connections necessitate some extra code, such as reinitialization of communications after faults, the final result is more resilient to error situations. RH specific failure mode analysis is covered in Section 5.

Another benefit of loose coupling between services is capability to have multiple copies running on more than one computer at once, providing redundant processing or hot backup capabilities. This can provide cost-efficient redundancy against computer hardware faults, especially for critical services. Services can also be hot swapped to enable better maintainability and non-stop reliability.

4. Remote handling control system

We have implemented an experimental RH control system to evaluate the proposed approach to fault detection and recovery. The RH control system operates an industrial manipulator (Fig. 2) manufactured by Comau in ITER relevant RH task scenarios. The
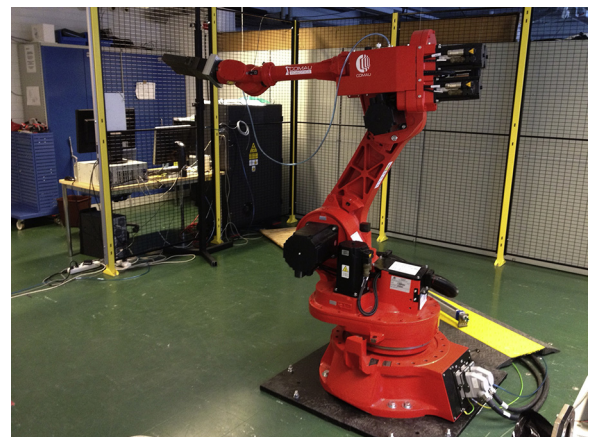

Fig. 2. Industrial manipulator used in experiments, equipped with a pneumatic gripper tool.

Table 1
RTSOA services on EC (service names in `Courier`).

Service	Service task description	Period T_i [ms]
`TrajectoryGenerator (TG)`	Generate a trajectory profile that the manipulator can follow from one point to another.	2
`C4G`	Send position and velocity reference to the manipulator control system at 500 Hz.	2
`C4GJointDataPub`	Publish manipulator joint position data to GSB.	10
`OmsCom`	Read OMS commands from GSB and pass them to the trajectory generator.	50

Table 2
RH control system service criticality & recovery. Stop = "stop manipulator movement", restart = "restart service".

Service name	Critical	Restore state	Failure action
`TG`	Y	N	Stop, restart.
`C4G`	Y	Y	Terminate TG, restart.
`C4GJointDataPub`	Y	Y	Stop, restart.
`OmsCom`	N	Y	Restart

system architectural model from Section 3.2 is applied to implement safe management of services in the RH control system to detect failures based on observing abnormal behavior. The RH control system is a non-trivial application that is used to test the fault detection and recovery in critical real-time systems. Communication frequency between master and slave controllers in similar teleoperation applications is typically 500–1000 Hz. The prototype uses 500 Hz communication loop between EC and Comau low-level servo controller over real-time Ethernet. Missing a communication deadline causes the low-level controller to engage an emergency stop and drop the communication link, necessitating a system restart.

Test setup includes EC, VR and OMS subsystems that are specified in the ITER RH control system architecture [2], in addition to the Comau's own low-level controller. VR is used to visualize robot position, although the operator also has direct visual contact with the manipulator in the test setup.

EC is a real-time system for operating RH equipment, i.e. the manipulator in this case. Our EC implementation is running real-time Linux with the RTSOA services and service manager. Services and loop timings are listed in Table 1.

VR capabilities are provided by IHA3D, Windows-based software developed at the Department of Intelligent Hydraulics and Automation (IHA) [13]. IHA3D provides a simulated virtual environment for the operator and can be used for virtual force generation in bilateral teleoperation.

OMS is a task planner subsystem used to support operation by planning, helping and instructing execution of RH procedures [13]. Procedures are complete sequences of manual actions and movements required to perform maintenance or testing operations. We used a web-server based OMS implementation which provides a browser-based GUI for the operator. Operator can use the OMS to send movement commands to the manipulator.

5. Fault recovery for RH system

5.1. Fault types

One way to categorize faults is permanent and transient faults [3]. Transient faults include aging-related faults (e.g. memory leaks), interaction faults, race conditions, attempts to exploit security vulnerabilities, resource leaks, bit flips, temporary device failures, etc. [3,10]. Transient faults can be difficult to find with testing because their activation may depend on complex timing, state and runtime environment conditions. This means that even critical systems developed with best practices can encounter them, making error confinement and recovery capabilities important. These faults may be temporarily solved by rejuvenation [14], i.e. shutting down the software item and restarting it. Although the root cause may not necessarily be removed by creating a fresh item, system

would typically be usable after restart. Transient faults are detected by service manager with resource limits, service status updates, heartbeat timeouts and monitoring of child process status.

Our implementation uses real-time message queues to send heartbeat signals from services to the service manager. Heartbeat signal is combined with the status update for the service. This functionality can be extended to the GSB, providing remote nodes awareness of service health. Resource limit monitoring is currently based on the proc file system, a feature in UNIX-like systems providing a method to access process data through a file-like structure. Alternatively system calls, such as getrusage, could also be utilized.

Permanent faults, caused, e.g. by faults in algorithms or control flow, persist after service restart. Determining if the fault is permanent or transient is based on error counting implemented by the service manager, i.e. limiting the number of restarts per service to prevent infinite restart loops that would otherwise be caused by former.

Severe permanent faults are typically detected in testing for commonly used features. If new permanent faults are encountered during operations, operator can avoid triggering conditions (if known) and recover RH equipment using reduced or alternative functionality for maintenance. For example, remote handling systems typically offer alternative control modes in the form of controllers and input modes (e.g. manual vs. OMS). If a permanent fault prevents use of OMS, operator can recover the equipment using manual mode. Loosely coupled service-based design supports this kind of robustness, as faults are isolated to services (e.g. `OmsCom`) and system can use alternative service configuration. Basically, a service failure is not necessarily a system failure.

5.2. Recovery from service failures

Next we describe services failures and recovery in the prototype RH system (see Table 2). Fault detection and recovery are implemented as previously described. Tests with the prototype system have shown that the service manager implementation is capable of detecting crashed services with heartbeat timeouts and process status. Crashed service is restarted or terminated according to the used configuration file. To be able to safely recover services, we need to understand roles of different services, as stated in [12], including identifying safety-critical services. For example, even if a non-critical service such as `OmsCom` experiences a permanent fault, system can retain partial operationality. VR and OMS are standalone applications that have separate software architectures and are presumed to have relevant fault tolerance mechanisms.

`Trajectory Generator (TG)` failures: steps for recovery process are outlined in Fig. 3. If a failure occurs during movement, robot needs to be stopped with a ramp because of large masses for remotely operated equipment, but this feature is typically implemented either in mechanical design with brakes or in the service responsible for robot movements (`C4G` service) as part of standard fault handling. Recovery of interrupted trajectory would be possible, but is not desirable in order to avoid sudden unwanted movements.

`C4G` failures: service failure stops all manipulator movements and is recovered to a safe state. Service manager must terminate
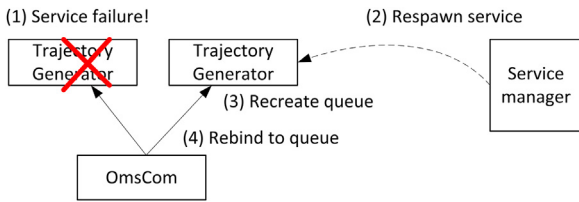
Fig. 3. Service recovering from failure; illustration simplified by showing only a single connected service. Solid arrows indicate the direction of data flow in queues.

also the TG service to avoid unintentional movements after restart. Service needs to recover state data for manipulator position, control mode and equipped tools. Position data can be recovered from sensors, other data can be recovered from GSB.

C4GJointDataPub failures: service can be used for VR visualization and haptic feedback, including virtual constraints and walls. System operation is possible without this service if operator has direct or video-based vision of the manipulator. However, since manipulator position data is used by support systems such as VR and collision detection, manipulator movement is stopped. Operation can resume safely after restart. State data restoration consists of reading joint status data from LSB queue.

OmsCom failures: service is non-critical in the sense that it is only used to initiate movement commands. In the case of permanent faults, manual operations with input device can be used for recovery. Command data is restored from GSB.

Compared to a non-critical application, such as a general-purpose OS in [10], first priority for a critical system is to guarantee safety of all operations. Fault in a critical service may compromise this goal so the system must move to a safe (fail-stop) state. In the case of a critical service, such as C4G, a fail-stop failure is preferred to undetected fault recovery. Therefore, critical services could be restarted to a *stopped* state. In the current implementation, services resume execution after restart (*start* command issued by service manager). In any case, services must be designed, implemented and tested to initialize to a safe state, even after recovery.

5.3. Recovery timing analysis

We analyze a situation where a failure occurs in the TG service. Failure is detected by the service manager and the service is restarted. Faulty service is presumed to report *error* status to the service manager (not shown in the figure) immediately after failure through heartbeat status updates that are sent during every service cycle. Cycle period T_i is 2 ms for C4G and TG services. Service manager uses $T_i = 1$ ms in the prototype. Worst-case jitter for periodic scheduling was tested on our previous study to have been around 6 µs, less than 1% of the RH service cycles [1].

To maintain correct operation, the C4G service must be able to read the updated trajectory values from TG during each period T_i. Therefore, to retain availability while recovering from service failure, combined time to detect failure (T_{detect}), time to restart ($T_{restart}$), and time to send and receive message (T_{com}) must be less than T_i plus possible spare time left from the previous cycle (T_{spare}).
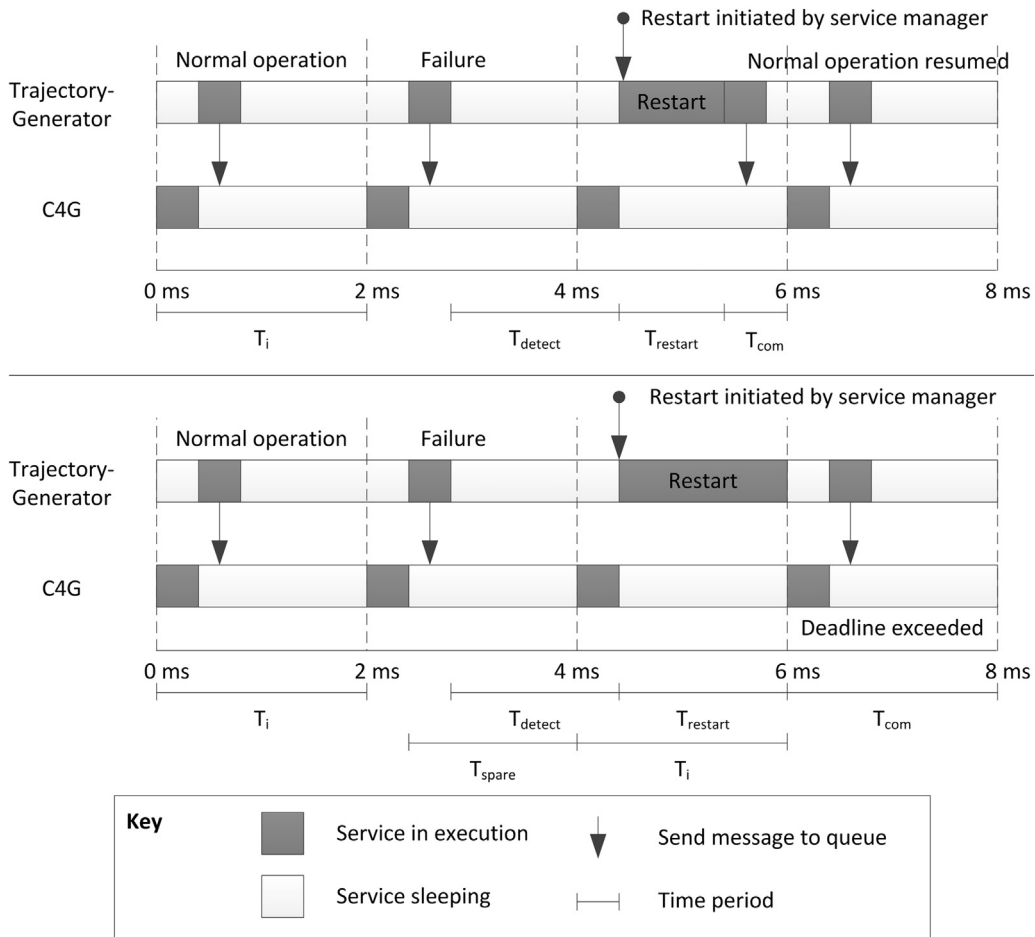

Fig. 4. Timing chart for service failure and restart. Upper scenario shows restart without exceeding communication deadlines and lower scenario restart with a missed deadline.

T_{spare} is included if the service has sent a correct message to C4G's LSB queue during the previous period, otherwise $T_{spare} = 0$:

$$T_{detect} + T_{restart} + T_{com} < T_i + T_{spare}$$

If true, operation can be resumed successfully. Otherwise C4G detects an exceeded communication deadline and the failure propagates. Fig. 4 illustrates both an optimal recovery situation and a "deadline exceeded" situation. C4G service has higher priority and is therefore executed first.

Successful recovery within time limits is largely dependent on fault detection time T_{detect} and restart time $T_{restart}$. Service manager cycle time is a trade-off point: if the service manager uses a fast cycle time ($<T_i$ of monitored services), it can react faster, but uses more CPU time. Alternatively, choosing a slow cycle time means that any service failure would automatically cause deadline miss ($T_{detect} > T_i$), forcing the system to a fail-safe state and causing a system failure.

5.4. Performance costs

Evaluation of performance costs of fault detection and recovery is challenging, since there is no comparable non-service based implementation available, making meaningful performance comparison problematic. However, other studies using similar approach with operating systems have estimated performance costs to be around 5–10% [10]. Taking into account the use of the service manager and communication queues to services, we estimate that the overall performance cost is around this order of magnitude. However, the exact number depends directly from the number of services, how often CPU and memory usage is checked and service manager task period T_i.

6. Conclusions

The decoupled architecture model – in this case the RTSOA – is one approach to managing challenges of complexity and scale. Based on the evaluation of our prototype implementation, it can support handling of transient faults and implementation of fault tolerance design patterns in critical real-time systems such as RH control systems. Although any architectural model cannot make the system automatically fault tolerant, it can provide tools for handling and mitigation of errors. A system based on a loosely coupled architecture can be in a safe state even after a service fault, without losing all system functionalities. The system is more robust to failures, providing mid-ground between binary "works" vs. "broken" options.

A large monolithic and tightly coupled application is difficult to verify, validate and maintain, whereas a service can be managed and verified individually, as long as the deployment environment is specified. Therefore, an approach based on loose coupling (e.g. service or component based) is recommended for ITER. Moreover, the RTSOA allows capabilities to be dynamically introduced to the system that were not initially planned. This is a major concern for a long-term research project such as ITER where the systems are likely to evolve to meet changing scientific and technical needs.

References

[1] P. Alho, J. Mattila, Real-time service-oriented architectures: a data-centric implementation for distributed and heterogeneous robotic system, in: 4th IFIP TC 10 International Embedded Systems Symposium, 2013, pp. 262–271.

[2] D. Hamilton, ITER Remote Handling Control System Design Handbook 2EGPEC v2. (2011) 3.

[3] A. Avizienis, J.-C. Laprie, B. Randell, C. Landwehr, Basic concepts and taxonomy of dependable and secure computing, Trans. Dependable Secure Comput. 1 (1) (2004) 11–33.

[4] T. Anderson, J. Knight, A framework for software fault tolerance in real-time systems, IEEE Trans. Softw. Eng. 3 (SE-9) (1983) 355–364.

[5] O. Gonzalez, H. Shrikumar, J.A. Stankovic, K. Ramamritham, Adaptive fault tolerance and graceful degradation under dynamic hard real-time scheduling, in: Proceedings of the 1st IEEE Real-Time Systems Symposium, 1997, pp. 79–89.

[6] E. Strunk, J. Knight, Dependability through assured reconfiguration in embedded system software, IEEE Trans. Dependable Secure Comput. 3 (3) (2006) 172–187.

[7] L. Sha, Using simplicity to control complexity, IEEE Softw. 4 (18) (2001) 20–28.

[8] X. Liu, Q. Wang, S. Gopalakrishnan, W. He, L. Sha, H. Ding, et al., ORTEGA: an efficient and flexible online fault tolerance architecture for real-time control systems, IEEE Trans. Ind. Inform. 4 (4) (2008) 213–224.

[9] B.U. Kim, Y. Al-Nashif, S. Fayssal, S. Hariri, M. Yousif, Anomaly-based fault detection in pervasive computing system, in: Proceedings of the 5th International Conference on Pervasive Services, 2008, pp. 147–156.

[10] J. Herder, Building a Dependable Operating System: Fault Tolerance in MINIX 3, Vrije Universiteit, Netherlands, 2010.

[11] P.M. Melliar-Smith, L.E. Moser, Progress in real-time fault tolerance, in: Proceedings of the 23rd IEEE International Symposium on Reliable Distributed Systems, 2004, pp. 109–111.

[12] P. Alho, J. Mattila, Breaking down the requirements: reliability in remote handling software, Fusion Eng. Des. 88 (9–10) (2013) 1912–1915.

[13] L. Aha, K. Salminen, A. Hahto, H. Saarinen, J. Mattila, M. Siuko, et al., DTP2 control room operator and remote handing operation designer responsibilities and information available to them, Fusion Eng. Des. 86 (9–11) (2011) 2078–2081.

[14] R. Hanmer, Software rejuvenation, in: 17th Conference on Pattern Languages of Programs, 2010.