FUJITSU

WHITE PAPER

Frequently Asked Questions on High Availability

Issue July 2012

Pages 11

Introduction

Nowadays, information technology (IT) services are often business-critical processes that contribute directly to the success of an enterprise. Functions such as e-business, enterprise resource planning (ERP) and supply chain management are placing demands on the information infrastructure that are growing at an astonishing rate. On the Internet, most business applications have to be available 24 hours a day, 7 days a week. High availability (HA) of the digital infrastructure is therefore more important than ever before for successful business activities.

The purpose of this document is to provide answers to various questions relating to high availability in general, high availability and cluster technology, the offering of Fujitsu for enhancing availability, and the HIPLEX AF (Highly Integrated system comPLEX Availability Facility) product family for the BS2000/OSD platform.

Contents

1 General questions on high availability

1.1 What is availability?

IEEE defines availability as the degree to which a system or a component is operational and accessible when required for use by an authorized user. In accordance with the usual formula for quantifying availability, availability is expressed in terms of a probability

$$\frac{\text{total operating time - downtime}}{\text{total operating time}} \times 100 \ (\%).$$

High availability is not a specific technology; it is rather an objective for which the solution needs to be tailored according to the particular situation of a company. A combination of strategies, technologies, training of employees, and different service levels is necessary in order to enhance the availability of the hardware, operating system, middleware, network and application(s) to a level that is acceptable to the customer.
In practice, a distinction is made between *planned* and *unplanned* downtimes.

1.2 What do the terms MTBF and MTTR mean?

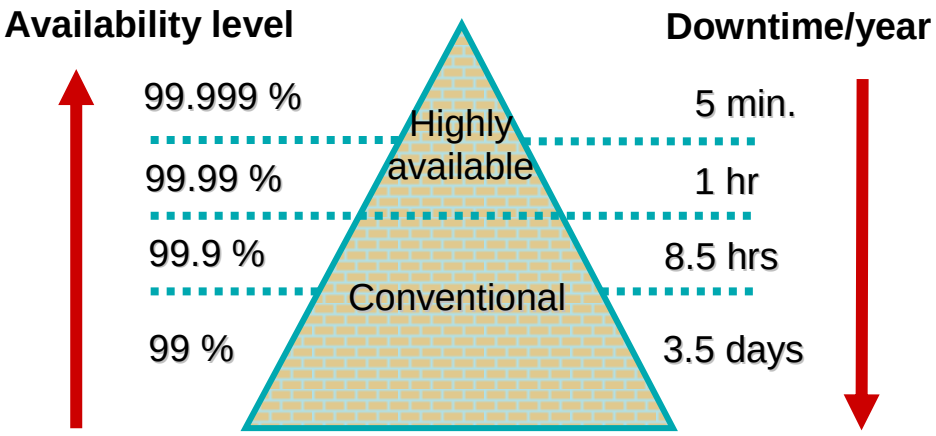
MTBF and MTTR are statistical expressions: for a system or a component, MTBF is defined as the **M**ean **T**ime **B**etween successive **F**ailures (or other types of outage), whereas MTTR is defined as the **M**ean **T**ime **T**o **R**epair such a failure and/or to recover from it.
Availability of a system or component can also be expressed in terms of MTBF and MTTR as

$$\frac{\text{MTBF}}{\text{MTBF + MTTR}} \times 100 \ (\%).$$

Note that when embedded in a large and complex system, the MTBF and MTTR values for a selected partial component may differ noticeably from the values for the very same component in an isolated environment.

1.3 Which downtimes are associated with the various levels of availability?

The amount of downtime accepted by customers depends significantly on their business needs. To identify this acceptable level, an in-depth analysis of the business processes must be carried out. This will influence the size and cost of a high-availability solution.
By using the formula mentioned in section 1.1, the following figures can be assumed:



1.4 How is availability to be seen in relation to reliability?

Reliability of software is one of the quality factors in software engineering and is defined as the extent to which a program can be expected to perform its intended function with the required precision within a specified period of time. Its rating formula is given by

$$1 - \frac{\text{no. errors occurred}}{\text{no. lines of code run through}}.$$

Reliability of hardware components is defined in an analogous manner.
It should be pointed out that availability percentages – in contrast to reliability figures – give no indication of the *frequency* of failures and of the *duration* of any single downtime.
For example, a system with availability of 99% over a whole year could fail four times a day for 3 minutes each time, which might be just within what can be tolerated. On the other hand, a single failure involving a continuous downtime of 84 hours – leading to the same availability percentage – would not be acceptable for most companies.

(Back to Table of Contents)

1.5 How is high availability to be seen in relation to disaster recovery?

Usually, high availability aims at circumventing disturbances of the production process (i.e. of the mission-critical applications) *within* a data center (for instance outages of *single* resources, or the introduction of new SW components in an IT system).
In contrast to this, **disaster recovery** denotes the safeguards needed to enable production operation to be resumed after partial or total destruction of a data center.
In the IT sector, a **disaster** usually refers to the failure of a data center due to power outage or destruction (through fire, flooding, explosion, earthquake, storm, sabotage, etc.).

1.6 How can the availability of a system be improved?

Fujitsu Siemens Computers offers solutions for enhancing system availability – besides other measures – by reducing the downtimes in the event of an occurrence of one of the events listed in sections 1.8 and 1.9.
For each platform, the solution essentially consists in introducing a high-availability cluster (see section 2.1) under a corresponding high-availability monitor. Among others, Fujitsu offers the following HA monitors in its portfolio:

Platform	HA Monitor
BS2000/OSD	HIPLEX AF
Linux	x10sure
MS Windows 2003 & 2008	x10sure

In addition, further organizational measures have to be taken (see the 'Business Continuity, High Availability and Disaster Recovery' set of slides). These include setting up an HA control center, running regular contingency practices, safeguarding the employees' expertise, enhancing the quality of service, and introducing an operations management model (preferably based on the ITIL de facto standard).
The availability of a system can further be improved by establishing a workload-balancing cluster (see section 2.16).

1.7 Can high availability be introduced gradually?

Yes. First of all, the redundant infrastructure must be established for each platform under consideration.
In general, high-availability monitors allow resources to be monitored in a flexible and granular way. The number and type of monitored resources may be changed, and the granularity of monitoring may also be decreased or increased.
These features allow a gradual introduction of high availability starting with a small number of applications and resources under the control of a high-availability monitor, which may be increased in subsequent steps.
Additional systems may also be added to the high-availability cluster.
The availability and the extent to which the data is up-to-date can also be increased gradually using functions of the operating systems, dedicated products, or special features of disk and tape storage subsystems (see section 1.10).
Comprehensive configurations comprising HA clusters on several platforms can be integrated into a monitoring management platform using the SNMP interface, for example (see chapter 3: references and links).
The organizational measures can also be taken gradually.

1.8 For which types of planned downtimes does Fujitsu offer a solution?

Fujitsu offers a solution for the following types of planned downtimes:
- Archiving, backing up or reorganizing data,
- Introducing, exchanging or upgrading hardware components (incl. maintenance),
- Introducing, exchanging or upgrading software components (in the operating system and/or application(s)), and/or
- Introducing software corrections (in the operating system and/or applications).

1.9 For which types of unplanned downtimes does Fujitsu offer a solution?

Fujitsu offers a solution for the following types of unplanned downtimes:
- A failure in a hardware component such as a CPU, a peripheral controller or device, or a data connection,
- A failure in the operating system (incl. middleware),
- A failure in an application,
- A failure in the data communication network,
- An operating error by the operator or system administrator, and/or
- The destruction of the entire data center ("disaster recovery").

(Back to Table of Contents)

1.10 How can the availability and extent to which the data is up-to-date be improved?

The availability and extent to which the online data is up-to-date can be improved by mirroring data either
- at a software level by using particular features of the operating system such as Dual Recording by Volume (DRV) in BS2000/OSD, or
- at a hardware level by using RAID disks and/or, for example, the SRDF or TimeFinder facility of Symmetrix, the SnapView or MirrorView facility of CLARiiON CX or the REC facility of ETERNUS DX400/8000 Disk Storage Systems.

A mechanism must exist, which makes all data available on all systems of the high-availability cluster.
After having been split off, a mirrored device can also be used for data backup on another system thereby avoiding a planned downtime in the production system.
To enable an application to bridge-over an unplanned downtime consistently, it is customary to use a transaction mechanism, so that the application can restart on a well-defined consistency point.

1.11 How can loss of data be avoided when an unplanned downtime occurs?

See section 1.10.

1.12 Can high availability be combined with disaster recovery?

Yes, provided the nodes in the HA cluster are located at sufficient distances from one another and remote data mirroring is used. A suitable means for implementing remote data mirroring at a hardware level is, for instance, using the MirrorView facility of CLARiiON CX, the SRDF facility of Symmetrix or the REC-facility of ETERNUS DX Disk Storage Systems.
In addition, some further organizational measures have to be taken. Besides the ones mentioned in section 1.6 (for details please see the IT Baseline Protection Manual of the “Bundesamt für Sicherheit in der Informationstechnik BSI (German Information Security Agency)” at http://www.bsi.de/gshb/index.htm), these include:
- Deploying contingency precautions,
- Preparing a contingency procedure handbook, and
- Running regular contingency practices.

2 Questions on high availability and cluster technology

2.1 What is a high-availability cluster?

A high-availability cluster consists of a number of systems with a common monitoring and failover infrastructure, which has no single point of failure. The resources of the system – databases, files, applications and devices, for example – may be made available on a cluster-wide basis. This enables the redundancies that exist in the cluster to be used throughout the cluster to avoid interruptions.
Applications which are switched to another system in case of a failure may run on all systems. Applications that are already running on this target system may then
- continue to run there (maybe with restricted performance), or
- be terminated (in case of minor availability requirements for them).

Several features smoothly cooperate in a high-availability cluster:
- Rapid, unambiguous detection of failed systems,
- Controlled and safe access to the cluster-wide resources from every system of the HA cluster,
- Automatic switching to another system in the cluster by an HA monitor in case of a failure, and
- Easy, central administration.

2.2 Is cluster technology necessary to achieve high availability?

With today's state-of-the-art technology, an availability of more than 99.99 % (corresponding to about 1 hour of downtime per year) can only be achieved with a cluster of systems (see section 2.1).

2.3 Which degree of application availability can be achieved in a high-availability cluster?

In special configurations up to 99.999 % can be achieved, corresponding to a downtime of 5 minutes per year ("five nines, five minutes"). The attainable value depends mainly on the time required for failover and thus on the actual application and configuration (see section 2.19).

2.4 Which types of high-availability cluster does Fujitsu support?

Fujitsu currently supports homogeneous high-availability clusters, which means that one and the same high-availability monitor monitors a cluster.
For HA monitors offered by Fujitsu see section 1.6.

(Back to Table of Contents)

2.5 What is the difference between a failover cluster and a high-availability cluster?

2.5.1 Failover cluster

A failover cluster consists of two or more active systems that normally run different applications. If the production system fails, the main application (or applications) and its/their resources are transferred to the standby system where recovery takes place. Since all systems in a failover cluster can be productively used in the normal case independently from each other, the failover concept represents a cost-efficient solution to the problem of failure management.
However, in a failover cluster it is entirely up to the operator to detect the failure, and to carry out the switching of the applications.

2.5.2 High-availability cluster

A high-availability cluster is far more elaborate than a failover cluster. In addition to the features described above, a high-availability cluster has a number of mutually independent and automatic monitoring functions for the connected systems, enabling unambiguous and immediate error detection. The automatic and rapid switching that takes place in case of a failure further improves the level of availability, thus eliminating the possibility of errors committed by the operator.
HIPLEX AF is an example of high-availability cluster.

2.6 What are the advantages of a high-availability cluster compared to a single system in the context of planned downtimes?

Hardware components (incl. maintenance) can be installed, exchanged or upgraded asynchronously on a standby system while the application is still running on the production system.
Equally, a new software configuration for the operating system and middleware (including version upgrades and correction updates) can be loaded and started asynchronously on a standby system while the application is still running on the production system.
In case of a problem in the operating system or the application after switchover of the application, the application can automatically be switched back to the previous production system.
The following figure schematically compares the procedures in a planned downtime with and without a failover to a standby system:

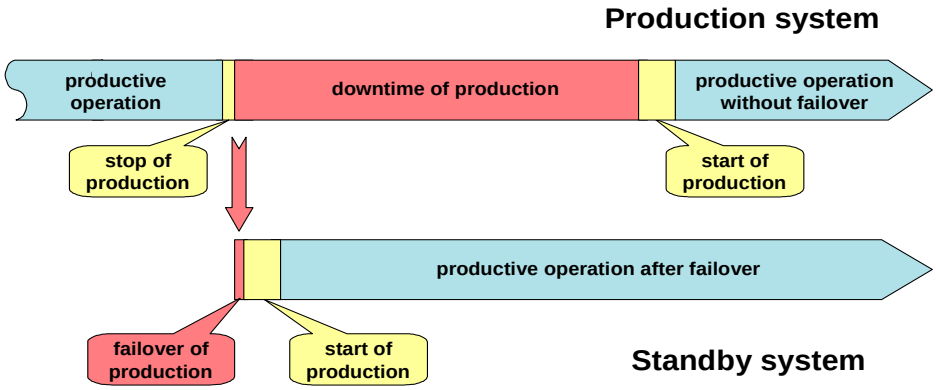


Figure 1: Planned downtime with and without a failover of the production to a standby system.

(Back to Table of Contents)

2.7 What are the advantages of a high-availability cluster compared to a single system in the context of unplanned downtimes?

A system failure is automatically and immediately detected.
In the event of a failure, the application(s) will automatically be switched over to a standby system, and the failed system can be repaired while the production is running on the standby system.
HIPLEX AF optionally tries to restart an application, which has failed, on the current node before carrying out a failover.
The following figure schematically compares the procedures in the event of a failure in a single system and an HA cluster:

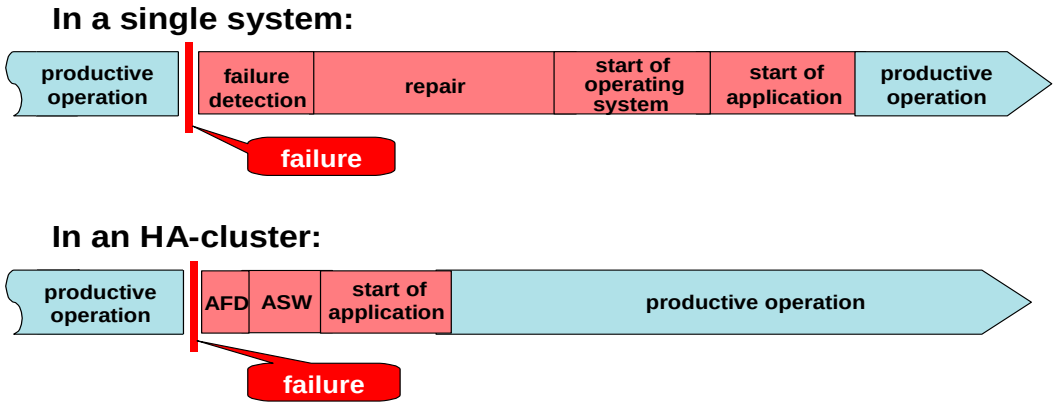


Figure 2: Comparison of actions in the event of a failure in a single system and an HA cluster.
Legend: AFD : Automatic failure detection
ASW: Automatic switch to another system.
The duration of AFD and ASW is limited compared with the duration of failure detection and repair in a single system, which is in principle unlimited.

2.8 Can a high-availability cluster also be established on one hardware system?

Yes, if the customer is happy with the availability of their hardware system, but would like to improve the availability of the software components. Several cluster nodes can be introduced by using different partitions. However, if the customer wants to increase the availability of both their software and hardware components, this solution is not appropriate.
In a BS2000/OSD environment, the VM2000 virtual machine system can also be used as the basis for the entire cluster (see section 4.2).

2.9 What is the maximum distance between two nodes in a high-availability cluster?

The limitations do not result from restrictions imposed by the high-availability monitor, but from the configuration and the applications running on the cluster. In most cases, remote data mirroring and shared data access are applied. The workable distances depend on the technology used (e.g. fibre channel) and the minimum accepted I/O throughput. Current costs of interconnection technology in conjunction with the speed of light usually limit the extension of clusters to approximately 10 km.

2.10 What is the difference between a hot and a cold standby system?

A *hot* standby system runs while it is in standby mode (with or without some applications being executed on it), whereas a *cold* standby system is not running while in standby mode and has to be started before an application can be switched to it.
With a cold standby solution, the customer therefore experiences a delay in service in comparison with a hot standby solution.

2.11 Do the standby systems have to be dimensioned exactly as the production system?

No, this is not necessary. Initially, the standby system(s) may be equipped with few resources. After a switchover has taken place, these resources can be reconfigured so that the standby system can run the business-critical applications at the requisite performance level.
Capacity on demand can also be provided as an option, which is useful in order to enhance the performance of a standby system after a failover has taken place.

2.12 Can single points of failure be avoided in a high-availability cluster?

Yes. All components of a high-availability cluster are installed redundantly to avoid any single point of failure.

2.13 Can applications also be monitored in a high-availability cluster?

Yes, of course. Monitoring of applications is the main objective in a high-availability cluster.

(Back to Table of Contents)

2.14 Which requirements must applications satisfy so they can be monitored in a high-availability cluster?

Most general-purpose applications, which run on a single system, are implemented in such a way as to allow them to run in a cluster in an unmodified manner. The main requirements for this are that they can be started and stopped under program control, that they can recover from failure situations, and that they provide a means to monitor their state.

2.15 Can distributed applications also be supported in a high-availability cluster?

Yes. Distributed applications are also supported in a high-availability cluster. Examples for such applications are the Oracle Parallel Server OPS, or Oracle RAC (Real Application Cluster).

2.16 Can dynamic workload balancing be established between several cluster nodes?

HIPLEX AF provides a command interface enabling the intentional switching of applications from one node to another (including their software environment). In order to be able to do this, in HIPLEX AF so-called "switch units" must have been introduced beforehand.
For distributed applications, see section 2.15.

2.17 Can a high-availability cluster react appropriately to changes in workload?

Yes. See section 2.16.

2.18 Must the user know on which node the application is currently running?

No. In HIPLEX AF the user can address an application by means of a unique node address, which is independent of the location where the application is currently running.

2.19 How much time is required to perform a failover in some sample configurations?

Generally, the following parameters significantly influence a failover time:
- The time necessary to detect the failure in case of an unplanned downtime,
- The performance of the CPUs involved,
- The time necessary to switch over the peripheral devices,
- The time necessary to unlock the peripheral devices in the event of an unplanned downtime,
- The time necessary to establish the execution environment on the target system; examples for this are print and file transfer jobs, which were still waiting in a system queue of the former production system,
- The time necessary to stop the application in the event of a planned downtime,
- The time necessary to restart the application, and
- The time necessary to perform the recovery of the application.

The time necessary for performing a failover therefore depends on the specific configuration and application.

2.20 What costs have to be considered when introducing a high-availability cluster?

When setting up a high-availability cluster, customers must take the following factors into account in their investment decisions:
- **Infrastructure costs**
  Infrastructure measures in the data center such as uninterruptible power supply, introduction of an operations management model, configuration updates, …
- **Product costs**
  Hardware (redundancy of the servers, devices, interconnections, etc.)
  Software (system version update, requisite and recommended products such as HIPLEX AF, HIPLEX MSCF, SDF-P, JV product family, where appropriate virtual machine system, etc.)
- **Service costs**
  Consulting costs, project costs for customer-specific implementation of the HA solution, etc.
- **Running costs**
  Service agreements, lease costs, maintenance of the cluster, etc.

However, please note: HA projects are usually performed with consolidation in mind.
Running costs are therefore reduced!

(Back to Table of Contents)

3 References and links

- Home page of Fujitsu on BS2000/OSD High Availability:
  http://ts.fujitsu.com/products/bs2000/concepts_solutions/high_availability.html
- Home page of Fujitsu for HIPLEX:
  http://ts.fujitsu.com/products/software/cluster_technology/hiplex.html
- Home page of Fujitsu for HIPLEX AF:
  http://ts.fujitsu.com/products/software/cluster_technology/hiplexaf.html
- HIPLEX AF – Manual:
  http://manuals.ts.fujitsu.com/index.php?l=en&id=1-2-3471
- VM2000 Virtual Machine System – Manual:
  http://manuals.ts.fujitsu.com/index.php?l=en&id=1-2-4969
- BS2000/OSD Management with SNMP:
  http://ts.fujitsu.com/products/software/networking/network_management/snmp.html

4 Questions/answers on HIPLEX AF

4.1 What is the functionality of HIPLEX AF?

An introduction can be found on the home page of Fujitsu for HIPLEX AF, especially within the data sheet linked there; a good description of the functionality of HIPLEX AF can be found in the section 3.1 “HIPLEX AF Basics” of the HIPLEX AF manual (see chapter 3 References and links)

4.2 What are the basic components in a HIPLEX AF cluster?

The basic components comprise hardware and system software components, as well as user procedures.

Hardware:

- Two or more servers (not necessarily with equal performance) in a configuration consisting of native systems, and/or systems on top of the VM2000 virtual machine system, or
- One server when using several VM2000 guest systems on the server

Coupling of these systems takes place by means of MSCF/BCAM communication links between the nodes and, maybe in addition, by means of a shared public volume set (SPVS).

System Software:

- HIPLEX AF and HIPLEX MSCF on each node of the cluster,
- BS2000/OSD operating system on each node of the cluster,
- VM2000 on every node of the cluster that is realized as a VM2000 guest system, especially in case of a virtual HIPLEX.

Virtual HIPLEX:

In a ***“virtual HIPLEX”***, the VM2000 virtual machine system is used as basis for the entire cluster. The systems on the various nodes are implemented as guest machines of VM2000.

Standard User Procedures:

Standard user procedures handle the failover of system components such as AVAS, openFT, SPOOL.

4.3 What other useful components exist for enriching a HIPLEX AF cluster?

Hardware components:

Symmetrix and CLARiiON CX disk storage subsystems for automatic remote data mirroring (SRDF or MirrorView).
On S series servers, the Global store GS can also be used in this environment to speed up access to shared data. Fujitsu uses the term "***parallel HIPLEX***" when the nodes in the cluster share a global store under control of XCS.

Software components:

SHC-OSD (support of SRDF in Symmetrix disk storage subsystems), DRV, HIPLEX MSCF with the XCS, XCS-TIME components

4.4 Which general requirements must be met for HIPLEX AF to be introduced?

A high-availability cluster based on HIPLEX AF consists of at least two different systems which are connected by HIPLEX MSCF, and which (in the current implementation) are connected to a common Shared Public Volume Set (SPVS). In addition, the subsystem JV (Job Variable) must be available.
If high availability is to be introduced on one server the VM2000 virtual machine system must be used.

4.5 Can different versions of BS2000/OSD and/or different versions of HIPLEX AF be installed on the nodes of a cluster?

The version of HIPLEX AF must be the same on each node of the cluster. For the BS2000/OSD versions, there are no restrictions except for the ones mentioned in section 4.6.

4.6 Which BS2000/OSD versions and which hardware models are supported?

All current BS2000/OSD versions, all current server hardware (S series, SX series and SQ series business servers) and the storage systems Symmetrix and CLARiiON CX are also supported by HIPLEX AF.

4.7 What is the maximum number of nodes in a HIPLEX AF cluster?

At present, the maximum number of nodes is 16 if the VM2000 virtual machine system is not used.
If the virtual machine system is used, for example for S models, the maximum number of nodes is restricted to 15 due to firmware limitations. Depending on the model of the machine, it may be further restricted to 7 (see VM2000 manual).

(Back to Table of Contents)

4.8 Can HIPLEX AF distinguish between an actual system failure and a bogus failure?

In the event of a bogus failure, the production system is still alive, even the applications might be untouched. Such a failure might therefore be ignored; an example for this may be a failure of each connection of the monitored system to the monitoring

system. The differentiation between actual system failure and a bogus failure is not possible in this case, but HIPLEX AF / HIPLEX MSCF override any automated failover and leave the decision to the operator.
HIPLEX AF uses HIPLEX MSCF to detect a failure in a monitored system. To make the information more reliable, HIPLEX MSCF uses different connection paths and different connection variants, which have to be simultaneously disturbed to assert that a failure occurred in the monitored system.

4.9 What is the influence on performance when a system is part of a HIPLEX AF cluster compared to a single system?

Monitoring the nodes in a HIPLEX AF cluster does not usually affect the system's overall performance.
A noticeable impairment of the system's performance only occurs in the phase when monitoring by HIPLEX AF is started.
In a virtual HIPLEX every virtual CPU of the production system that is declared in a guest system under normal circumstances leads to a performance reduction of at most 1%.

4.10 What is the purpose of SDF procedures?

SDF procedures provide a menu-driven interface for adjusting a configuration to specific customer needs. They offer a sound functionality by providing interfaces to practically all system functions.

(Back to Table of Contents)

Published by department:

Barbara Stadler
Phone: ++49 (0)89 62060 1978
Fax: ++49 (0)89 62060 329 1978
Barbara.Stadler@ts.fujitsu.com
ts.fujitsu.com