A novel approach to fault tolerant multichannel networks designing problems

Mirosław Hajder\textsuperscript{1*}, Paweł Dymora\textsuperscript{1†}

\textsuperscript{1}Rzeszów University of Technology, Faculty of Electrical and Computer Engineering, Department of Distributed Systems, ul. Wincentego Pola 2, 35-959 Rzeszów, Poland

Abstract – This work presents solution of a bus interconnection network set designing task on the base of a hypergraph model. In order to do this the interconnection network is presented as a multipartite hypergraph. A system with virtual bus connections functioning in an environment of common physical channel was analyzed, which is characteristic of the networks based on the WDM technology. The mathematical reliability model was proposed for two modes of system functioning: with redundancy of communication subsystem and division of communication load. As solution estimation criteria the expected changes of processing efficiency changes were used as also a communication delay change criteria and system reliability criteria. The designing task solution is searched in a Pareto set composed of Pareto optima. The selection procedure of a specific solution in the case of its equivalency in relation to a vector goal function was presented.

1 Introduction

In the last decade widely observed development in the field of information transmission was an advanced integration of computer networks and telecommunication systems. Telecommunication took over the transmission methods and means of computer science which are based on the massive use of computer techniques. It also took over computer tools used in traffic management. On the other hand, computer science does the services reserved so far for telecommunications (especially voice and picture communication as well as motion picture transmission). Communication technologies

\textsuperscript{*}Miroslaw.Hajder@gmail.com

\textsuperscript{†}Pawel.Dymora@prz.edu.pl
A novel approach to fault tolerant multichannel networks

such as: ATM – Asynchronous Transfer Mode, FR – Frame Relay, MPLS – Multi Protocol Label Switching are equally addressed to both network types and may become a tool for their effective integration. Nowadays, from the communication point of view a distributed system can have a heterogeneous character, i.e. equally based on the resources of these two network types.

The other important feature of current distributed systems is a variety of offered services. Nowadays in the same network different, often incompatible communication services (e.g. isochronous and synchronous transfers) exist. A novel approach to the problem of QoS (Quality of Service), which is based on the dynamic allocation of independent communication channels for users or services present in the network is needed. In particular, it refers to multimedia services done in real time or critical infrastructure system (CIS) of user’s functioning. The analyzed systems are also characterized by high dynamics of exploitation parameters change. Network load is not actually of static character, it dynamically changes, thus it prevents from designing and implementation of the network satisfying even short-term user’s requirements. Besides, in the networks there exists traffic having a burst character (bursty traffic) whose bursty coefficient, at specific moments, makes proper network functioning impossible. Efficient solution of most presented problems is an assurance of flexible connection reconfiguration, best on the logical level without necessity of hardware architecture modification. In this way, connections may be dynamically adopted to an actual network traffic pattern. As a result, creating connections for distributed systems we go away from static designing towards incremental and evolutional methods.

The effective connections reconfiguration methods may arise due to implementation of modern communication technologies, especially these which allow their accomplishment on the logical level and additionally improving the level of physical communication channels use. The particular example of this situation is WDM (Wavelength Division Multiplexing) technology, which on the basis of optical communication environment allows building multichannel networks with direct connections also as multipoint network. Admittedly, it appears to be very interesting to build multichannel bus networks with shared logical bus and the ability of dynamical change of the bus set to which end the user is connected, owing to this it is possible to adjust network architecture to traffic patterns existing in it. However, applying such an architecture a solution of specific designing is needed, which for obvious reasons should be characterized by accepted time and memory complexity which is difficult to acheive due to the combinatorial task character.

The above designing task may be subsumed to a wide class of system building configuration tasks. This task is mostly divided into three main subtasks: a. system components selection; b. its allocation; c. defining the connections between them. In the previous works the system components selection subtask was solved, among others, with the use of implemented methods, e.g.: searching the shortest path [1, 2], block backpack [2, 3], clustering in multipartite graph [4], multipartite clique [5], morphological analysis. In turn, to solve components allocation subtask, generally
different variants of assignment task are used nowadays\[2, 6, 7, 8\]. To define connections between the system components, generally the agglomeration \[9, 10\] and solution methods of optimal hierarchy building task \[11\] are used.

The subject of the presented research is implementation and application of hypergraph theory methods and algorithms in order to solve the above designing task. Its aim is creating of the designing task solution method which is characterized by the polynomial time complexity.

2 Connection architecture and system reliability model

Let us consider a distributed system shown in Fig. 1 composed of \(K_N\) computation nodes, where each of them has \(K_l\) tunable send-receive elements and \(K_B\) communication buses. Buses \(B_i (i = 1, \ldots, K_B)\) have a logical character and may be realized on the basis of the wavelength division multiplexing methods in a single physical bus \(D_F\). Any node \(N_i\) logical channels connecting to a physical bus \(B_F\) is achieved using physical connection channels \(l\) and bus channels separator \(C\).

![Fig. 1. Generalized bus computation system architecture.](image)

The analyzed system connection architecture may be dynamically reconfigured by tuning send-receive elements. In the case of a traffic pattern change the send-receive elements change the wavelength and actually connect to other logical bus. The system may work in two modes: \textbf{a.} with a communication subsystem redundancy; \textbf{b.} with communication load division. To estimate characteristics of the system working in one of these three modes, a probabilistically-combinatorial model was proposed, especially for estimating its reliability \(R\). In this model, the set: send-receive device \((\approx)\), physical connection link \((l)\) and physical channel separator \((C)\) is treated as one single connection device. Let \(p_{io}\) – be is a functioning probability of a computation node send-receive device, \(p_l\) – a physical connection channel functioning probability;
Let us consider the bus distributed system reliability with complete connections (each computation node is connected to each logical bus) and processing equal rights computation nodes, which is the most generalized example of the system of this class.

The probability \( p_{we}(k_{we}) \) of connection sets functioning which assures connection to a bus channel not less than \( k_{we}^{\min} \) computation nodes from the general amount of \( K_N \) is expressed as:

\[
p_{we}(k_{we}) = p_{fk} \sum_{i=k_{we}^{\min}}^{K_N} C_k^i p_{ku}^i (1 - p_{ku})^{k_{we} - i}.
\] (1)

Let \( K_B \) denote the number of logical buses, \( p_{we}^{sp} \) – the computation node functioning probability. Then, using (1), according to the proposed condition, for a system in a redundancy mode, the reliability \( R \) is expressed as:

\[
R = \sum_{j=k_{we}^{\min}}^{K_N} C_j^{K_N} (p_{we}^{sp})^j (1 - p_{we}^{sp})^{K_N - j} \sum_{k=1}^{K_B} C_k^{K_B} p_{we}(j)^k (1 - p_{we}(j))^{K_B - k}.
\]

Let us consider a computation system with equal computation nodes which works in a communication load assignment mode. Let \( W(k_{we}, k_m, \sigma) \) denote the number of system functioning states consisting of \( k_{we} \) computation nodes connected by \( k_m \) bus channels with occurrence of \( \sigma \) denies of send-receive elements. In order to obtain the number of system functioning states \( W(k_{we}, k_m, \sigma) \), in the case of failure elements appearance, the methods of inclusion-and-exclusion was proposed. The number \( H_1(\sigma) \) of the whole system failure states in the case of deny not less than one minimal system cut with equal computation nodes may be expressed as:

\[
H_1(\sigma) = K_N C_{(K_N - 1)K_B}^{\sigma - K_B} + C_{K_N}^{\sigma - K_B} \sum_{\alpha=1}^{K_B-1} C_{K_B}^{\alpha}.
\]

Then the value \( W \) may be expressed as:

\[
W(K_{we}, k_m, \sigma) = C_{k_{we}k_m}^{\sigma} + \sum_{i=1}^{i_{\sigma}} (-1)^i H_i(\sigma),
\]

is the summary number of system states for \( \sigma \) denies of send-receive elements; \( i_{\sigma} \) – the maximal number of minimal cuts included in estimation; \( H_i(\sigma) \) – the number of computation system failure states with denies of send-receive elements not less than \( i \) minimal cuts. The coherence assurance probability \( P_{ku}(k_{we}, k_m, \sigma) \) of \( k_{wu} \) nodes with the use of \( k_m \) buses, in the case of occurrence \( \sigma \) denies can be expressed as:

\[
P_{ku}(k_{we}, k_m, \sigma) = W(k_{we}, k_m, \sigma) p_{ku}^{k_{we}k_m - \sigma} (1 - p_{ku})^\sigma.
\] (2)
Using expression (2), the coherence assurance probability \( P_{ku}(k_w, k_m) \) of \( k_w \) computation nodes with the use of \( k_m \) buses, in the case of occurrence \( \sigma \) denies can be expressed as:

\[
P_{ku}(k_w, k_m) = \sum_{\sigma=0}^{(k_w-1)/k_m} P_{ku}(k_w, k_m, \sigma).
\]

Using expression (3), let us define coherence assurance probability \( P_{aku}(k_w) \) of \( k_w \) computation nodes:

\[
P_{aku}(k_w) = \sum_{l=k_m}^{K_B} C_l^K_B p_{f_k}^l (1 - p_{f_k})^{K_B-l} P_{ku}(k_w, l),
\]

where: \( K_m^{\text{min}} \) is the minimal required number of buses necessary for capacity assuring. Then, the computing system reliability is equal to:

\[
R = \sum_{n=k_w}^{K_N} C^n_{K_N} (p_{w}^{sp})^n (1 - p_{w}^{sp})^{K_N-n} P_{aku}(n).
\]

For a client-server computing system in the redundancy mode we can define a bus communication channel reliability probability as \( P_{ks}(K_K, K_S) \) where through correct send-receive elements there are connected not fewer than \( k_k^{\text{min}} \) clients from the whole client number \( K_K \) and \( k_s^{\text{min}} \) servers from the whole server quantity \( K_S \):

\[
P_{ks}(K_K, K_S) = p_{f_k} \sum_{i=k_k^{\text{min}}}^{K_K} \sum_{j=k_s^{\text{min}}}^{K_S} C_i^{K_K} p_{ku}^i (1 - p_{ku})^{K_K-i} C_j^{K_S} p_{ku}^j (1 - p_{ku})^{K_S-j}.
\]

Let \( p_k^{sp} \) and \( p_s^{sp} \) be the client and server nodes activity probability, respectively. Then, using expression (4), the reliability \( R \) may be expressed as:

\[
R = \sum_{m=k_k^{\text{min}}}^{K_K} C_m^{K_K} (p_k^{sp})^m (1 - p_k^{sp})^{K_K-m} \sum_{n=k_s^{\text{min}}}^{K_S} C_n^{K_S} (p_s^{sp})^n (1 - p_s^{sp})^{K_S-n} \\
\sum_{l=1}^{K_m} C_l^{K_m} (P_{ks}(K_K, K_S))^l (1 - P_{ks}(K_K, K_S))^{K_m-l}.
\]

Let us consider the client-server system reliability where the system has complete connections between computing elements and communication subsystem working in a load division mode. Let \( W(k_s, k_k, k_m, \sigma) \) be a number of correct system states composed of \( k_s \) servers, \( k_k \) clients, \( k_m \) buses, with occurrence of \( \sigma \) denies of send-receive elements. For the client-server system, the number of \( H_1(\sigma) \) failures states for not fewer than one minimal cut can be defined as:

\[
H_1(\sigma) = (k_k + k_s)C_{k_m}^{\sigma-k_m} (k_k + k_s)C_{k_m}^{\sigma-k_m} = C_{k_m}^{\sigma-k_m} (k_k + k_s + k_k k_s \sum_{\alpha=1}^{k_k k_s} C_{k_m}^{\alpha}).
\]
and the number of correct functioning states as:

\[ W(k_k, k_s, k_m, \sigma) = C_{k_m(k_k+k_s)} + \sum_{i=1}^{i=\sigma} (-1)^i H_i(\sigma), \]

where: \( C_{k_m(k_k+k_s)} \) - the general number of computing system states with occurrence of \( \sigma \) denies. The coherence assurance probability \( P_{ku}(k_s, k_k, k_m, \sigma) \) of \( k_s \) servers and \( k_k \) clients with the use of \( k_m \) buses in the case of occurrence \( \sigma \) send-receive elements denies can be expressed as:

\[ P_{ku}(k_s, k_k, k_m, \sigma) = W(k_s, k_k, k_m, \sigma) p_{ku}^{(k_s+k_k)k_m-\sigma}(1-p_{ku})^{\sigma}. \] (5)

The coherence assurance probability \( P_{ku}(k_s, k_k, k_m) \) of \( k_s \) servers and \( k_k \) clients with the use of \( k_m \) buses in the case of send-receive elements deny appearance is expressed as:

\[ P_{ku}(k_s, k_k, k_m) = \sum_{s=0}^{s=(k_s+k_k)k_m} P_{ku}(k_s, k_k, k_m, \sigma), \] (6)

and the coherence probability \( P_{ku}(k_s, k_k) \) of \( k_s \) servers and \( k_k \) clients can be defined as:

\[ P_{ku}(k_s, k_k) = \sum_{k_m=1}^{K_m} C_{k_m}^{k_k} p_{fu}^{k_k} (1-p_{fu})^{K_m-k_m} P_{ku}(k_s, k_k, k_m). \] (7)

Using expressions (5), (6) and (7) the searched reliability \( R \) may be expressed as:

\[ R = \sum_{k=k_{min}}^{K_k} C_{K_k}^{k_k} (p_{ku}^{sp})^k (1-p_{ku}^{sp})^{K_k-k} \sum_{s=k_{min}}^{K_s} C_{K_s}^{s} (p_{ku}^{sp})^s (1-p_{ku}^{sp})^{K_s-s} P_{ku}(s,k). \]

3 Main definitions

Let us consider a hypergraph \( H = (V, E) \) with a node set \( V = \{v\} \) and an edge set \( E = \{e\} \) which represents subsets of the \( V \) set e.g. \( e \subseteq V \). Hypergraph \( H \) is called \( k \)-regular, if its each edge \( e \in E \) consists of \( k \) nodes. Then the hypergraph \( H \) is called \( l \)-partite if its node set is divided into \( l \) subsets. In a \( l \)-partite hypergraph a node set \( V \) is divided into \( l \) subsets \( V_1, V_2, \ldots, V_l \), in such a manner that all edge nodes \( e = (v_1, v_2, \ldots, v_l) \in E \) are parts of other graph regions, e.g. \( v_i \in V_i \), where: \( i=1,\ldots,l \). To denote \( l \)-partite hypergraphs we use a notation \( H = (V_1, V_2, \ldots, V_l) \).

Let us consider the \( l \)-partite hypergraph denoted as \( H = (V_1, V_2, \ldots, V_l) \). In this graph, its part \( a = (V_1^A, \ldots, V_l^A) \) is called a star, for \( i = 1,\ldots,l \) and \( V_i^A \subseteq V_i \), in which any two edges \( e_1, e_2 \in E_A \) cover in one and the same node \( v \in V_i^A \) and don not cover in any other \( v \in V_i^A \). It means that the set cardinality \( V_1^A \) is equal to \( 1 \), and the node \( v \in V_i^A \) is called a star center. We distinguish simple and complex stars. If any edge pair \( e_1, e_2 \in E_A \) covers only in one node \( v \in V_1^A \), then such a star is called a simple star. In another case a star is called a complex star. A number of star edges is called a star degree. For a star edge \( e = (v_1, v_2, \ldots, v_l) \in E \) nodes \( v_1 \) and \( v_l \) are called end nodes. Therefore, nodes \( v_2, \ldots, v_{l-1} \) are called inner nodes. The node
sets of a graph part $V_2, \ldots, V_{l-1}$ are composed of an empty pair of disjoint sets $V_i(v_j)$, $v_j \in V_j$, where: $i = 2, \ldots, l - 1$, $j = i + 1$.

For a hypergraph $H = (V, E)$ its subhypergraph $H_1 = (W, U)$ is called a hypergraph in which the node set $W$ is a subset of a node set $V$ of the hypergraph $H$ e.g. $W \subseteq V$ and an edge set $U$ is a subset of an edge set $E$ of the hypergraph $H$, whereas, if $(x, y) \in E$ and $x, y \in W$, then $(x, y) \in U$. The hypergraph nodes set is called a hypergraph coherence component, in which for its any two elements there exists a path from a node belonging to this set to any other node apart from this node set. If in a subhypergraph $H_1 = (V_1, E_1)$ of the hypergraph $H$ each coherence component is a star with a center in a node $v \in V_1$, than $H_1$ is called a hypergraph $H$ star cover.

4 Designing task and its solution

Designing task solution consists in searching a connection architecture which ensures its exploitation parameters maximization or minimization such as: communication delay, communication medium access errors being a consequence of its business, computation efficiency of processing nodes and others. Such a task may be solved by searching a star cover of at least three-partite graph. Let us consider such a task.

**Input data.** As an input data of a designing task we use: 1) $B = \{b\}$ – a set of logical communication buses selected in a physical communication channel with the use of a wavelength multiplexing method WDM [12, 13]; 2) $F = \{f\}$ – a set of access protocols, according to which a logical communication bus can function [12]; 3) $N = \{n\}$ – a set of system clients using bus communication channels. Clients are divided into groups $d \in D$ including their communication requirements where $D = \{d\}$ is a communication types set. The set $D$ elements may have the following values: $d = 0$ – the service of streams sensitive to transmission errors and delays; $d = 1$ – the service of streams sensitive to transmission delays; $d = 2$ – the service of streams sensitive to transmission errors; $d = 3$ – the service of streams not sensitive to below factors.

**Designing task definition.** Each of processing nodes $n \in N$ should have an assigned set $M$ of $b$ buses, which is a subset of the set $B$, e.g. $M \subseteq B$, each of those will be functioning on the basis of one access protocol $f \in F$.

**Mathematical model.** The designing task is solved iterationally. In each step for each of the processing nodes one bus $b \in B$ is searched serving for a node communication services with the use of an access protocol $f \in F$. In order to avoid multiple nod joins to the same bus, each step from the set of buses available for clients each of those to which client is connected is excluded.

The mathematical model is based on the three-partite hypergraph $H = (V, E) = (X, Y, Z, E)$. Bus-es from the set $B = \{b\}$ have equivalent nodes $x$ of its first part ($x \in X$). Each of them (contemporary each of logical buses) a label $\eta(x)$ is assigned defining bus transmission characteristics, in the simplest case, the number of clients which this bus may serve.
The elements $f$ from the set $F$ of bus access protocols correspond to nodes $y$ of the second part of the hypergraph $H$ ($y \in Y$), and for the $n$ elements of $N$ processing nodes nodes $z$ of the third part of the hypergraph ($z \in Z$) corresponds. The edge set $E = \{e\}$ has all node trios $(x, y, z)$ such as $x \in X$, $y \in Y$, $z \in Z$. Only such edges are allowed for which the selected bus $b_i \in X$ may serve client $n_j \in Z_3$ with the use of communication protocol $f_i \in Y_2$. The set $E = \{e\}$ of all edges is defined as a set of all allowed node trios $e = (x, y, z)$. Including the value of parameter $n(x)$, for $x \in X$ in the hypergraph $H = (V, E) = (X, Y, Z, E)$ the allowed step solution of the designing task is any of its subhypergraphs $\beta = (V_\beta, E_\beta)$, for $V_\beta \subseteq V$ and $E_\beta \subseteq E$, which each coherence component represents a simple star with the degree $\eta(x)$ and the center in a node $x \in X$. As $S = S(H) = \{s\}$ let us denote a set of all allowed solutions of the hypergraph $H$ star covering task.

Each edge $e \in E$ of the hypergraph $H = (V, E)$ has assigned three weights, defining the following solution characteristics:

1. $\omega(e) = \varphi(x, y, z)$ – the expected change of client processing efficiency in a system in which a client $z$ is communicationally served by a bus $x$, in which communication protocol $y$ is used. To estimate processing efficiency we may use characteristics proposed in [14, 15, 16]. The characteristics presented there were modified in order to reflect the efficiency change which is an effect of connection architecture changes. Taking into consideration a bus character of the interconnection network the basic efficiency characteristic is a node number for which a communication network is accessible [17].
2. $\xi(e) = \varphi(x, y, z)$ – the expected change of communication delay on a client demand for the above conditions. The level of delay changes may be defined on the base of stochastic model which uses the method presented in [18];
3. $\psi(e) = \varphi(x, y, z)$ – the expected change of the error level while realizing client communication demand for a saved condition in step 1. To estimate error level changes the method presented in Section 2 was used.

**Designing task solution.** The solution estimation will have a multicriteria character. The proposed criteria set has obviously sample character and its selection depends on the designer needs especially related to future exploitation of the interconnection network. For a considered case these are three functions presented below.

Let us consider a set $A = \{a\}$ of acceptable solutions of designing task. For each of them we define the following characteristics which estimate the solution quality:

1. The computation efficiency criteria: $\Phi_1(a) = \max_{a \in A} \min_{e \in E_a} \omega(e)$, where: $E_a$ is a set of edges of the hypergraph $H$ belonging to the solution $a$. Using this criterion we aim to maximize the minimal level of system efficiency (computation or communication) as the whole.
2. The communication delay criteria: \( \Phi_2(a) = \min \sum_{e \in E_n} \xi(e) \), which ensures searching interconnection network with the minimal summary delay. For a system with different level of nodes priority the value \( \xi(e) \) of expected change of communication delay is calibrated with the use of node priority.

3. The reliability criteria: \( \Phi_3(a) = \max \sum_{e \in E_n} \psi(e) \). Similarly to the case of communication delay criteria \( \Phi_2 \), this criterion ensures a network architecture searching for which a summary reliability is maximal.

The possibilities of the above method do not limit applying summarizing criteria in a form of \textit{min} or \textit{max}. To estimate a solution quality it is allowed to use freely parameter convolution methods, including those involving weights of individual subparameters.

The above partial criteria are connected by the goal function expressed as \( \Phi(a) = (\Phi_1(a), \Phi_2(a), \Phi_3(a)) \). The multicriteria goal function \( \Phi(a) \) is used which defines in a set \( A \) of accepted solutions a Pareto set \( A^p \) composed of Pareto optima \( a^p \). Let us assume that two solutions \( a_1, a_2 \in A \) of a vector goal function \( \Phi(a) \) are equivalent. Then from the Pareto set \( A^p \) a full set of alternatives \( A^A \) is selected which is generally a maximal system of different vectored Pareto optima.

5 Research, results and further work

The approach presented in this work was used to create a designing methodology of multichannel bus connections in order to communicationally serve client-server computing systems.

Existing methodologies are concentrated on a defined capacity level assurance of the whole computing system, skipping its reliability parameters. In this work the presented methodology searches an optimal solution in relation to the multicriteria goal function where one criterion is reliability. Depending on a partial criterion relationship manner and constraints set the proposed methodology allows also to specify a connection architecture characterizing: a. maximal reliability with defined: minimal efficiency and maximal communication delay in an interconnection network; b. minimal communication delay with a constraint in minimal reliability and efficiency; c. maximal computing efficiency with a defined acceptable level of reliability and delay. Also there was tested a solution in which a designing goal is communication buses load balancing. For each of searched solutions maximal building costs were limited. We analyzed networks with complete and partial connections as also flat and hierarchical.

Simulation research of the obtained architectures for a client-server model was based on the methodology presented in [19] and presented in the Fig. 2. In both cases the optimal structure in relation to a vector criterion was searched. The following notation was used: \( K_s, K_k \) - the server and client number respectively; \( \omega \) – the request submitting probability; \( \omega_p \) – the probability of a request submitted to a preferred server; \( \omega_k \) – the probability of a request submitted to a server of the same group; \( \omega_o \) – the probability of a request submitted to any server of other groups.
Further research will concentrate on searching the effective methods of multipartite graph coverage searched with a simple star which allowing using graphs of any partitions in a design-ing process, thus allows to introduce further designing criteria.
References