A quantitative analysis of the performance of functional distributed architectures

S.Zaporojan

Abstract

This paper is concerned with the functional distributed architectures (FDA) performance. In this work a quantitative analysis of FDA performance has been carried out. The carried out analysis makes it possible to determine the limits for the effective usage of FDA structures. The formulae for the speedup of the computations at the FDA structures have been obtained. The formulae for determining the maximum number of executive processors which allows us to achieve the greatest computation’s speedup at the given FDA structure have also been obtained. It is shown that use of the executive processors with communication autonomy allows us to introduce the additional levels of the instruction’s interpretation, due to which the bus traffic is reduced. A quantitative evaluation of this fact is presented in the paper.

KEYWORDS: functional distributed architecture, control and executive processor, cyclic interleaved computations, speedup, macrooperation, one/two-level instruction’s interpretation.

1 Introduction

The purpose of this paper is to investigate the problem of the performance of FDA. We consider that the analysed FDA consists of two subsystems: control and executive [1,2]. It is supposed that the control subsystem contains a control processor (CP). The executive subsystem consists of one or more executive processors (EP). It is also considered that the communication between processors is performed via a common bus to which a system memory (SM) is connected.
The functional distributed architectures have been investigated in [1]. In this work a formal description of the FDA functioning based on the distribution between the control and executive processor of four phases of the instruction’s execution is given. These phases are: fetching of the instructions and input operands from system memory, processing of the input operands on the executive processor as well as writing down the results. In conclusion the principles of FDA organisation are proposed: operational, process, and communication autonomy of the executive processor. Using the proposed principles 16 typical structures have been presented.

It is obvious that to increase the speedup of computations the processors of FDA must operate simultaneously. That is, the computations with interleaving must be used. The overlapped computations are used in a special processor architecture, called the Interleaved Array Processor (IAP) [3]. The last work shows that a greatest speedup is achieved when the cyclic computations with overlapping are organised. The similar computations have been organised at a FDA based on a control processor and a FFT processor [4]. An important problem of how many processors can be used effectively when the computations are organised with overlapping was investigated in [3,5]. In [3] a quantitative analysis of the performance of matrix multiplication by the IAP is carried out. It is supposed that the data are transmitted via the two buses. A general analysis of the performance of the FDA is presented in [5].

The contribution of this paper is to provide a quantitative analysis of the FDA performance that takes into account the characteristics of control and executive subsystems, of common bus and executed algorithms. In section 2 some basic concepts such as basic macrooperation, one-level and two-level instruction’s interpretation as well as the notation are introduced. We are also presenting in this section the structure of the analysed system and an algorithm of computations based on a round robin scheduling technique. Sections 3 and 4 present a quantitative analysis of the FDA performance. The boundaries of the efficient usage of the given functional distributed architecture are determined taking into account the characteristics of control and executive processors, of common bus and executed algorithms. The application of the
A quantitative analysis of the performance of...

communication autonomy allows us to introduce the additional levels of the instruction’s interpretation, due to which the bus traffic is reduced. Therefore, it is possible to increase the computation’s speedup. In section 4 a quantitative evaluation of the above fact is obtained.

2 Basic concepts

Our aim is to provide a quantitative analysis of the FDA performance when the cyclic computations with overlapping are used. Let us consider the FDA shown in Fig.1.

![Diagram](image)

Fig.1. The structure of the analysed system

We shall suppose that the structure of the analysed system consists of one control and P executive processors. Communication between processors is performed via common bus to which a system memory is connected. We assume that there is enough system memory for any of executed algorithm. We will also consider that the cycle time of system memory is equal to the common bus access time. Besides, we suppose that the executive processor consists of an arithmetical unit and a control one as well as of four registers. Two registers are used as a source of operands and other two are used as a destination. The
executive processor structure can also contain a communication unit and possibly a local memory.

When considering the solution of a fixed problem, we let $T_{CP}$ denote the time required to execute the problem using only the control processor. The time required to solve the task on the FDA is $T_{FDA}$. Then the speedup of computations on the given FDA is defined as

$$S = \frac{T_{CP}}{T_{FDA}}.$$ 

There is known [6,7] the inequality, called Amdahl’s Law,

$$S \leq \frac{1}{\varphi},$$

where $\varphi$ is a positive “essentially serial” fraction of a computation. This inequality shows that the speedup is bounded when we increase the number of processors for a fixed problem. In practice we want to execute algorithms with various characteristics.

Let us suppose that an algorithm must be executed by the analysed FDA using $W = K + M$ instructions. It is supposed that $K$ and $M$ represent the number of instructions and macroinstructions that should be executed by the control and executive processors respectively. A macroinstruction we shall call one or more macrooperations with input operand vectors $X_1, \ldots, X_r$ of equal length $l$. If $Y_1, \ldots, Y_q$ are result vectors, then the macrooperation ($MOp$) is defined as follows,

$$(Y_1[j], \ldots, Y_q[j]) = MOp(X_1[j], \ldots, X_r[j]), \quad (1)$$

for $j = 1$ to $l$.

We will consider that it is possible to perform both basic macrooperations and the macrooperations of the first level. We will suppose that the executive processor is capable to execute any basic macrooperations, described as follows,

$$y_1 = MOp(x_1); \quad y_1 = MOp(x_1, x_2); \quad (y_1, y_2) = MOp(x_1, x_2). \quad (2)$$

If the basic macrooperations can be only performed on the given FDA structure then one-level interpretation of the macroinstruction is
used. If it is possible to perform the first-level macrooperations via basic macrooperations of the given FDA structure then a two-level interpretation of the macroinstruction is used. The computational power of basic macrooperations can be modified when the executive processor structure is changed. Hence, the computational power of the first-level macrooperations can be also changed.

Each macrooperation can be executed on the executive processor either consequently or using a round robin scheduling technique. Obviously one would expect a greater calculation’s speedup in the second case. We assume that it is possible to organise cyclic calculations at the analysed FDA. Then we shall consider the following. First, we suppose that input operands are stored in the system memory. Any final or intermediate result produced by the executive processor must be written back into the memory. Second, we shall consider the following algorithm of computations:

1. For $j = 1$ to $l$, the control processor fetches the input operands from $SM$ and forms a command packet consisting of the input operands and an $MOp$ code for executive processor.

2. These packets are sent to the executive processors through the common bus using a round robin scheduling technique.

3. When an executive processor receives a command packet it performs the macrooperation, designated by the $MOp$ code, with the input operands that are contained in the packet. When the macrooperation is finished, the executive processor forms an output packet containing the results.

4. The output packet containing results is sent to the system memory through the common bus by the control processor.

Mention that the input operands fetching and the results storing can be performed by the executive processor when it has communication autonomy. In this case the control processor only forms the command packets and sends these packets to the executive processors.

In order to analyse the performance of the FDA, the following notation is used:
$t_A$ is the common bus access time. This is the time required to transfer 
one item of data between the system memory and an executive 
processors.

$t_{MOV}$ is the transfer time of one item of data by communication unit 
via the common bus. It is supposed that $t_{MOV} = t_A$.

$t_{CP}$ is the time required to perform the simple arithmetical operations 
or the transfer of one item of data by the control processor mul-
tiples of $t_A : t_{CP} = at_A$.

$T_F$ is the input operand’s packet fetch time in multiples of $t_{CP}$ or 
t$_{MOV}$.

$f$ is the number of elements in an input operand packet.

$T_{COM}$ is total time to transfer a command packet to an executive pro-
cessor.

$n$ is the number of elements in a command packet.

$N$ is the number of command packets.

$T_S$ is the output result packet download time in multiples of $t_{CP}$ or 
t$_{MOV}$.

$m$ is the number of elements in an output result packet.

$t_{OP}$ is the computation time of basic macrooperations on the arith-
metical unit in multiples of $t_A : t_{OP} = ct_A$.

$t_{EXEC}$ is the command packet processing time on the executive pro-
cessor.

$t_{ILVD}$ is the interleaved time. This is the time when the work of control 
and executive processors can be overlapped.

$T_{MOP}$ is the computation time of basic macrooperations on the control 
processor.
A quantitative analysis of the performance of...

The macrooperation execution on the executive processor may be represented as follows,

\[
\frac{(Y_0 - Y_1)}{M_0} \rightarrow \mu \rightarrow \frac{(X_0 - X_1)}{N_0} \rightarrow t_{mac}
\]

where \( \mu = 1 \) when the basic macrooperations are only performed. When a macroinstruction is performed using a hierarchical interpretation mechanism [2], \( \mu \) represents the number of basic macrooperations that must be performed when interpreting a macroinstruction.

The number of processors that can be used effectively when the computations are organised with overlapping and using a round robin scheduling technique is determined [5] by,

\[
P = \max \left\{ \left\lfloor \frac{t_{EXEC}^i}{t_{LVD}^i} \right\rfloor \right\},
\]

where \( t_{EXEC}^i \) is the command packet processing time for the \( i \)-th macrooperation and \( t_{LVD}^i \) is the overlapped time when the \( i \)-th macrooperation is performed. The braces mean rounding of the enclosed value to the smallest integer greater than the initial number.

We will consider the question of whether it is possible to attain a desirable speedup of computations at the given FDA. In particular, the structure that contains the arithmetical unit and the communication unit will be used as an executive processor. Thus, the executive processor has operational and communication autonomy. Therefore, each executive processor is capable of fetching and executing basic and first-level macrooperations from system memory and can also store the
results back into the memory. Below the boundaries of the efficient usage of the given functional distributed architecture are determined taking into account the characteristics of FDA and executed algorithms. We first consider the case when the algorithm is executed using the one-level interpretation of macroinstructions. Then, the case when the algorithm is executed using the two-level interpretation of macroinstructions is investigated.

3 Analysis of the one-level interpretation

This section will investigate the performance of computations at the given FDA when the macroinstructions are interpreted via basic macrooperations described by (2), i.e. \( \mu = 1 \). Obviously, the arithmetical and communication units of the executive processor structure can operate either consequently or simultaneously. The first structure represents a serial-based executive processor. In the second case we have a parallel-based executive processor. Subsection 3.1 investigates the performance of computations on the FDA with serial-based executive processors and subsection 3.2 describes the performance of computations on the FDA with parallel-based executive processors respectively.

3.1 Computations with the serial-based executive processor

Let the arithmetical and communication units of the executive processor can only operate consequently. In this case the control processor forms the command packet containing the macrooperation code and the address of input operands in the memory. Thus, we may consider that. The formed command packet is then sent via the common bus to the executive processor. When the executive processor receives the command packet, the input operands are fetched from system memory by the communication unit of this processor. After that the arithmetical unit performs the macrooperation, designated by the MOp code that is contained in the command packet. The results produced by the arithmetical unit are then written into the memory by the communica-
A quantitative analysis of the performance of...
where 1 \leq i \leq N. From equation (8) it follows that \( P^S_{\text{max}} = 1 \) if
\[
t^i_{OP} \leq T^i_{COM}
\]
for any \( i \). The command packet processing time can be written in multiples of \( t_A \) as
\[
t^i_{EXEC} = (f^i + c^i + m^i)t_A,
\]
and the overlapped time,
\[
t^i_{ILVD} = (n^i a + m^i + f^i)t_A.
\]
Then (8) may be written as
\[
P^S_{\text{max}} = \max \left\{ \left[ \frac{f^i + c^i + m^i}{n^i a + m^i + f^i} \right] \right\}.
\]
When one and the same macrooperation is executed, i.e. \( i = 1 \), or when the macrooperations have identical \( T^{S_i}_{COM}, T^i_S, T^i_F, t^i_{OP} \) times for all \( i \), then
\[
P^S_{\text{max}} = \left[ \frac{T_F + t_{OP} + T_S}{T_{COM} + T_S + T_F} \right] = \left[ \frac{f + c + m}{n a + m + f} \right].
\]
The time required to execute the algorithm by the FDA may be written as
\[
T_{FDA} = T^K_{CP} + T_{ESS},
\]
where \( T_{ESS} \) is the time required to execute \( N \) macrooperations by the executive subsystem. This time is determined as follows:
\[
T_{ESS} = T_{COM} + T_F + (N - 1)t_{ILVD} + t_{OP} + T_S = Nt_{ILVD} + t_{OP}.
\]
The time required to execute \( K \) instructions by the control processor is
\[
T^K_{CP} = K t_{CP}.
\]
The time required to execute the algorithm using only the control processor
\[
T_{CP} = K t_{CP} + N(T_F + T_{MOP} + T_S),
\]
124
A quantitative analysis of the performance of...

with

\[ T_F = f_{CP}, \quad T_{MOP} = z_{CP}, \quad T_S = m_{CP}. \] (17)

Using the above presented relations the speedup of computations at the FDA with serial-based executive processors can be written as

\[
S = \frac{T_{CP}}{T_{FDA}} = \frac{t_{CP}(K + N(f + z + m))}{K t_{CP} + N(m t_{CP} + m t_{MOV} + f t_{MOV}) + t_{OP}}. \] (18)

Ignoring the time \( t_{OP} \) and taking into account that \( t_{CP} = a t_A \) and \( t_{MOV} = t_A \) we have

\[
S = \frac{a t_A(K + N(f + z + m))}{K a t_A + N t_A(na + m + f)}.
\] (19)

Let us introduce the following notation: \( K = \alpha N \), where \( \alpha \geq 0 \). Then the speedup may be written as

\[
S = \frac{a(\alpha + f + z + m)}{a \alpha + na + m + f}.
\] (20)

From (20) it follows that the speedup will be achieved when

\[
z > n + \frac{(1 - a)(m + f)}{a}. \] (21)

Using (17) the inequality (21) may be written as

\[
T_{MOP} > \left[ n + \frac{(1 - a)(m + f)}{a} \right] t_{CP}. \] (22)

Thus, the inequality (22) shows when the speedup of computations is available at the FDA with serial-based executive processors. The right part of this inequality may be negative what means that we should have achieved the speedup even if the computation time of basic macrooperation on the control processor had been equal to zero.

Obviously when solving a problem with \( N \gg K \), i.e. when \( \alpha = 0 \), a boundary speedup \( S_b \) will be achieved on the given FDA. Then from (20) it follows that

\[
S_b = \frac{a(f + z + m)}{na + m + f}.
\] (23)

125
For $a \to \infty$, i.e. when the control processor is very slow, from (20) it follows that
\[
\lim_{a \to \infty} S = \frac{\alpha + f + z + m}{\alpha + n}.
\] (24)
The formula (24) allows us to determine the possible maximum speedup of fixed computation on the FDA with serial-based executive processors.

3.2 Computations with the parallel-based executive processor

This subsection investigates the performance of computations at the FDA with parallel-based executive processors. Suppose that the arithmetical and communication units of the executive processor can operate simultaneously. Then the input operands processing on the arithmetical unit and the fetching of the next input operands by the communication unit (if the next command packet has been received) can be executed simultaneously. After the current processing is finished the arithmetical unit can process the next operands and the communication unit can execute the storing of the last result into the system memory.

Fig.3 shows the timing diagram when executing the algorithm at the FDA with parallel-based executive processors. It can be seen that the total overlapped time is given as
\[
t_{ILVD} = T_{COM} + T_f + T_s,
\] (25)
i.e. this time is identical with the overlapped time given by the formula (4). On the other hand, the command packet processing time is given as
\[
t_{EXEC} = t_{OP}.
\] (26)

Hence the maximum number of executive processors is determined by,
\[
P_{max}^P = \max \left\{ \left[ \frac{t_{OP}^i}{T_{COM}^i + T_f^i + T_s^i} \right] \right\},
\] (27)
A quantitative analysis of the performance of...

Fig. 3. The timing diagram

or, analogously to (11),

$$P^P_{\text{max}} = \max \left\{ \left[ \frac{c^i}{n^i a + f^i + m^i} \right] \right\}.$$  \hspace{1cm} (28)

For $i = 1$, or when the macrooperations have identical $T^i_{COM}, T^i_S, T^i_F, t^i_{OP}$ times for all $i$, we can write

$$P^P_{\text{max}} = \left[ \frac{t_{OP}}{T^i_{COM} + T^i_F + T^i_S} \right] = \left[ \frac{c}{na + f + m} \right].$$  \hspace{1cm} (29)

We would expect a reduction of the number of parallel-based executive processors to achieve the same speedup comparatively with the case when the computations are organised with the serial-based executive processors. The analysis of the diagrams shown in Fig.2 and Fig.3 as well as the comparison of the formulae (8) and (27) confirm the fact. In particular, dividing (12) and (29) we have the following theoretical estimation

$$\frac{P^S_{\text{max}}}{P^P_{\text{max}}} = 1 + \frac{T^i_F + T^i_S}{t_{OP}} = 1 + \frac{f + m}{c}.$$  \hspace{1cm} (30)

Practically the number of executive processors may be identical.

From the diagram shown in Fig.3 it follows that the time required to execute $N$ macrooperations by the executive subsystem with parallel-based executive processors is given as

$$T^i_{ESS} = Nt_{ILVD} + (t_{OP} - T^i_S) + T^i_S = Nt_{ILVD} + t_{OP}.$$  \hspace{1cm} (31)
It is seen that the obtained formula is identical with (14). Therefore, the formulae of the computation’s speedup obtained in previous subsection are true when executing the algorithm on the FDA with parallel-based executive processors.

4 Analysis of the two-level interpretation

We will now investigate the case when the algorithm is executed by the analysed FDA using the two-level interpretation of macroinstructions. Let us suppose that the algorithm is executed by the first-level macrooperations described by (1). Each first-level macrooperation is interpreted via the \( \mu \) basic macrooperations described by (2). The control processor forms the command packet that must contain the length \( l \) of input vector \( X \) and the address of input operands in the memory as well as the length \( r \) of the first-level macrooperation. Besides, the command packet can contain the address of the result to store it back into the memory. The formed command packet containing as a minimum \( n = 3 \) elements is then sent to the executive processor. It interprets the first-level macrooperation code executing \( \mu \) basic macrooperations. When processing the command packet the input operands are fetched from the system memory and the result produced by the executive processor is written back into the memory for each of \( \mu \) basic macrooperations. The two-level computations at the FDA with serial-based executive processors are illustrated by timing diagram shown in Fig.4.

The use of the two-level interpretation allows to exclude formation and transmission of the command packets when executing each of \( \mu \) basic macrooperations. Obviously the exclusion of the above mentioned operations permits us to reduce the bus traffic. Below an estimation of the mentioned fact is presented.

From the diagram in Fig.4 it is seen that the interpretation time \( t_{EXECL}^{IL} \) of the first-level macrooperation is defined as,

\[
t_{EXECL}^{IL} = \mu t_{EXECL}^{OL},
\]

where \( t_{EXECL}^{OL} \) is the processing time of the basic macrooperation. The total overlapped time when executing the first-level macrooperation is
A quantitative analysis of the performance of...

Fig. 4. The timing diagram
defined by

\[ t_{ILVD}^L = T_{COM}^{IL} + \mu(T_S + T_F). \]  \hfill (33)

Hence, in particular, the maximum number of executive processors will be calculated by

\[
P_{\text{max}} = \left[ \frac{t_{\text{EXEC}}^{IL}}{t_{ILVD}^L} \right] = \left[ \frac{\mu_{\text{EXEC}}^{QL}}{T_{COM}^{IL} + \mu(T_S + T_F)} \right] = \left[ \frac{t_{\text{EXEC}}^{QL}}{(T_{COM}^{IL} \mu) + T_S + T_F} \right]. \]  \hfill (34)

The maximum number of serial-based executive processors is determined as

\[
P_{\text{max}}^S = \left[ \frac{T_F + t_{\text{OP}} + T_S}{(T_{COM}^{IL} \mu) + T_S + T_F} \right] = \left[ \frac{f + c + m}{(\mu^{IL}a \mu) + m + f} \right]. \]  \hfill (35)

When the parallel-based executive processors are used we have \( t_{\text{EXEC}}^{QL} = t_{\text{OP}} \). Therefore,

\[
P_{\text{max}}^P = \left[ \frac{t_{\text{OP}}}{(T_{COM}^{IL} \mu) + T_S + T_F} \right] = \left[ \frac{c}{(\mu^{IL}a \mu) + m + f} \right]. \]  \hfill (36)

It is easy to observe that the formulae (12) and (29) are obtained from the formulae (35) and (36) respectively in the case \( \mu = 1 \), i.e. when the one-level interpretation is used. It is also easy to see that equation (30) is true when the two-level interpretation is used. A possible difference of the number of the command packet elements is necessary to be taken into account only when the one-level and two-level interpretation are used.

From equation (34) it follows the increase of the maximum number of executive processors when \( \mu \) is increased. That is, the use of the two-level interpretation allows to increase the implemented concurrency of the FDA. It is important to mention that when the three-level interpretation will be used then the increase of the maximum number of executive processors is very little. This assertion can be easily demonstrated using expressions similar to equations (32) and (33).
From the diagram shown in Fig.4 it follows that the time required to execute \( N \) first-level macrooperations by the executive subsystem is given as
\[
T_{\text{ESS}}^{\text{IL}} = T_{\text{COM}}^{\text{IL}} + T_F + (N - 1)t_{\text{ILVD}}^{\text{IL}} + \mu t_{\text{OP}} + T_S. \tag{37}
\]
Taking into account equation (33) the last expression may be written as
\[
T_{\text{ESS}}^{\text{IL}} = NT_{\text{COM}}^{\text{IL}} + (T_S + T_F)[\mu(N - 1) + 1] + \mu t_{\text{OP}}. \tag{38}
\]
When \( \mu = 1 \) the formulae (14) and (31) are obtained from equation (38).

The speedup \( S^I \) of the executive subsystem based on the two-level interpretation over the executive subsystem based on the one-level interpretation is the ratio of the times required to execute the macrooperations using the one-level interpretation and the two-level interpretation as follows,
\[
S^I = \frac{T_{\text{ESS}}^{\text{OL}}}{T_{\text{ESS}}^{\text{IL}}}, \tag{39}
\]
where the time required to execute \( \mu N \) basic macrooperations using the one-level interpretation is given as
\[
T_{\text{ESS}}^{\text{OL}} = N\mu \left( T_{\text{COM}}^{\text{OL}} + T_S + T_F \right) + t_{\text{OP}}
\]
\[
= \left[ N\mu(n^{\text{OL}}a + m + f) + c \right] t_A, \tag{40}
\]
and the time required to execute \( N \) first-level macrooperations using the two-level interpretation is
\[
T_{\text{ESS}}^{\text{IL}} = NT_{\text{COM}}^{\text{IL}} + (T_S + T_F)[\mu(N - 1) + 1] + \mu t_{\text{OP}}
\]
\[
= \left( Nn^{\text{IL}}a + (m + f)[\mu(N - 1) + 1] + \mu c \right) t_A. \tag{41}
\]
Then, we have
\[
S^I = \frac{N\mu(n^{\text{OL}}a + m + f) + c}{Nn^{\text{IL}}a + (m + f)[\mu(N - 1) + 1] + \mu c}. \tag{42}
\]
5 Conclusions

This paper has investigated the problem of how must the computations be organised so that to achieve a greatest speedup at the functional distributed architectures. The use of communication autonomy allows us to introduce the additional levels of instruction interpretation, due to which the bus traffic can be reduced. Therefore, it is possible to increase the number of executive processors. It is mentioned that when using the three-level interpretation the reduction of the bus traffic is very little. Therefore, from this point of view, the computations should be organised using the one-level and two-level interpretation. Using equation (42) an estimation of the speedup of the executive subsystem based on the two-level interpretation over the executive subsystem based on the one-level interpretation can be obtained. The executive subsystem uses $P_{max}$ processors in both cases.

References


A quantitative analysis of the performance of...


S.Zaporojan, Received 13 January, 1996
Computer Science Department,
Technical University of Moldova,
Ștefan cel Mare 168,
2012, Chișinău, Republic of Moldova,
phone: (+373.2) 497016

133