FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

A catalog of classifying characteristics for massively parallel computers

Tilmann Bönniger*, Rüdiger Esser, Dietrich Krekel*

KFA-ZAM-IB-9405

Februar 1994
(Stand 28.02.94)

(*) Regionales Rechenzentrum an der Universität zu Köln, Robert-Koch-Str. 10, 50931 Köln

To be published in the proceedings of HPCN Europe, April 18–20, 1994

A catalog of classifying characteristics for massively parallel computers

Tilmann Bönniger[1], Rüdiger Esser[2], Dietrich Krekel[1]

[1] Regionales Rechenzentrum an der Universität zu Köln, Robert-Koch-Str. 10, D-50931 Köln, E-mail: {Boenniger,Krekel}@rrz.Uni-Koeln.DE
[2] Zentralinstitut für Angewandte Mathematik, Forschungszentrum Jülich GmbH, D-52425 Jülich, E-mail: R.Esser@KFA-Juelich.DE

Abstract. In order to facilitate an application-oriented assessment of high-performance massively parallel computing systems, a catalog of about 350 classifying characteristics concerning the architecture and software environment of such systems has been compiled. The data required for the catalog allow a rather complete and homogeneous description of a massively parallel system to be established. This article contains an overview of the catalog and the hardware model on which it is based.

1 Introduction

Since a new generation of high-performance massively parallel systems (MP systems) has recently become commercially available and claims to be an effective alternative to traditional vector supercomputers, application developers as well as computing centers have to decide whether this new technology can serve their needs. In order to facilitate the assessment of MP systems, we established a catalog of classifying characteristics and prepared a comparative description of some specific MP systems on the basis of the catalog [2], similar to earlier work for vector supercomputers [1]. The data required for the catalog allow a rather complete and homogeneous description of a massively parallel system to be established. In contrast to the data sheets provided by the manufactures, the catalog facilitates the comparison of MP systems even if they have different architectures. What is more, gathering the data required for an MP system makes it possible to understand this system much better. In the process of selecting an MP system, this collection of data complements performance measurements with benchmark programs from the intended application area.

This paper contains an overview of the catalog and of the hardware model forming its basis. In another paper [3], the data required for the catalog have been gathered for three commercially available MP systems: Intel Paragon XP/S, Kendall Square KSR1, and Thinking Machines CM-5. That article also includes textual descriptions of the three MP systems based on the classifying characteristics as well as figures showing their overall architecture, processing nodes, vector units and memory hierarchy.

2 The hardware model

The catalog is based on a generic hardware model (Figure 1) that most of today's commercial MP systems seem to follow: a three-stage hierarchy of *processing nodes*, *clusters* of nodes, and the entire *MP system*, in which every stage has its *internal communication* network and facilities for communicating with the communication system of the next higher stage (*external communication*). At every stage, there may exist a *memory*. There are usually different types of nodes including *special nodes* supporting the operating system and Input/Output.

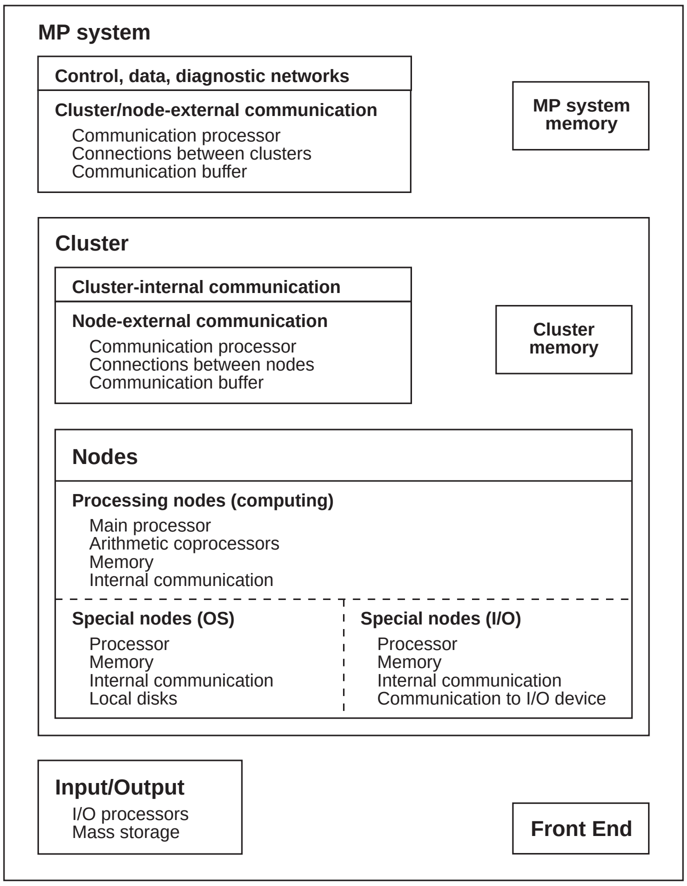

Fig. 1. Generic hardware model of the MP system architecture

3 Catalog of classifying characteristics

The catalog is meant to be used for an application-oriented description of available high-performance massively parallel computer systems. It consists of approximately 350 items which are arranged in 14 groups (a complete listing of all items [2,3] can be found in the appendix):

0. Marketing information
1. General characteristics of the MP system
2. Processing nodes
3. Special nodes
4. Cluster
5. MP system
6. Front end
7. Input/Output
8. Operating system
9. Compilers
10. Parallel programming concepts, tools, and libraries
11. Application software
12. Configurations
13. Performance

The first group *marketing information* helps to gain some insight into the role of the MP system division and its products within a company. Strong market-selection mechanisms have taught us well to consider economic aspects. The group *general characteristics of the MP system* (Table 1) allows a general classification and a rough description of an MP system to be made.

Table 1: General characteristics of the MP system

MP system	**Processing node**
global classification	node architecture
scaling range (sold systems)	processing elements
MP system architecture	clock frequency
single-stage / multi-stage (clustered) architecture	node-internal communication
node / cluster types	node-external communication
max. number of nodes / clusters	node memory
memory concept	. real / virtual (page size)
max. memory sizes	. address length
max. filesize	. address space
globally accessible diskspace	. size
typical access	**Input/Output**
connection with front end	concept
external networks	I/O bandwidth
Cluster	**Software**
cluster architecture	operating system
nodes	. MP system
intra-cluster connection	. node
shared memory size per cluster	file systems
extended memory	languages
I/O	programming models
	parallel CASE tools

The following groups give a detailed description of an MP system and its components in a bottom-up fashion. The group *processing nodes* (Table 2) refers to the functionality and the performance of a single node as part of an MP system. At the node level, a comparison with workstation functionality and performance is possible. For example, the nodes become more and more complex in order to keep pace with high performance workstations, thus complicating the task of the compiler. A trade-off between off-the-shelf and proprietary chips has to be considered. Apart from general processing nodes, MP systems contain *special nodes* implementing specific functions. The experience of the manufacturers with MP systems is reflected here, especially with respect to the services for which these nodes are provided and how they are integrated in the system.

Table 2: Processing nodes

Designation by manufacturer	error handling
Functionality	**Node-internal communication**
Architecture	intra-node connections
components	. type
clock frequency	. number and width
Main processor	. latency
technology	. bandwidth
chip type	**Node-external communication**
chips per processor / processor on a chip	communication processor
cache	. type (number)
registers and buffers	connections between nodes
integer performance (peak)	. type (number)
floating point performance (peak)	. latency
Data formats	.. best case (packet size)
classification of data representation	.. worst case (packet size)
length	. bandwidth node-to-node
classification of arithmetic	.. best case (packet size)
Arithmetic coprocessors	.. worst case (packet size)
technology	. bandwidth (one node to all other nodes)
chip type	.. best case (packet size)
functionality	.. worst case (packet size)
. integer performance	communication buffer
. floating point performance	number of independent communication operations
local memory	communication without interrupt of main processing
registers and buffers	**Hardware for performance measuring**
Node memory	**Additional special hardware**
size	**Possible handicaps and bottlenecks**
chip type, cycle time and size	
latency	
memory bandwidth	

It is not obvious whether the *cluster* as a level between the nodes and the MP system as a whole has to be considered. For example, the processor rings of the KSR system can be viewed as clusters. Other systems, e.g. the PARSYTEC GC system, fit well into this concept. There seems to be a trend towards lean hierarchies at least from an architectural point of view. The next group of classifying characteristics, *MP system*, is the most essential part of the catalog.

Apart from communication architecture, its implementation is essential for system performance. Here, too, the difference between an MP system and networked workstations becomes evident. Implemented connectivity, employed communication techniques, additional special hardware, e.g. concerning fault tolerance, and the backplane's functionality, are important system concepts and components, especially when it comes to the support of high bandwidth, low latency communication and real scalability. Therefore it is not surprising that detailed information on these topics is difficult to obtain.

For future MP systems the importance of the classifying characteristics concerning the *front end* might further decline because front ends tend to be integrated into the systems as interactive nodes. Available MP systems differ considerably in functionality, connectivity, compatibility of usage and system administration. *Input/Output* has been the poor cousin of MP systems. I/O system performance has not kept pace with the advances in processor and memory speeds. Well-known problem areas include I/O cache placement, interference between normal message and I/O traffic, parallelism extended to the file systems [4]. As MP systems will be used in real-time multimedia or transaction processing applications, I/O performance and system integration have to be enhanced and this will be reflected in future versions of the catalog, too.

Concerning software, 4 groups of characteristics have been established. The group *operating system* describes the functionality of the operating system from a user and data center point of view. None of the operating systems of current commercial MP systems is a truely distributed operating system according to the characteristic features presented by Tanenbaum [5], but the difference between an MP system and e.g. a network of workstations that use a common file system is evident. MP systems can be well distinguished in what and how operating system functionality (e.g. virtual resources, system administration) is implemented, the user interface generally being UNIX.

In the group *compilers* the functionality of available compilers, e.g. FORTRAN and C, is described especially with respect to parallelization and vectorization constructs. As exploiting special hardware of nodes may be effective, assemblers are included as well. The characteristics included in the group *parallel programming concepts, tools, and libraries* (Table 3) are important even from a management point of view, because the ease of use as well as the cost involved in getting an application to run optimally on the MP system determine user acceptance. Tools should support all stages of porting an application: initial port to a single processor, performance optimization on this processor, parallelization, performance optimization for the MP system. A general-purpose MP system should support the standard MIMD programming models:
- message-passing (including SPMD and manager/worker programming styles)
- data parallel (especially High Performance Fortran)
- shared memory (uniform global address space)

Table 3: Parallel programming concepts, tools, and libraries

Support for programming models	**Other tools**
SIMD parallel model	**Native communication libraries**
Data parallel model	message passing
MIMD shared memory model	informational routines
Message passing model	global synchronization
Program development tools	global operations
name (type; functionality; runs on)	remote message passing
Debugging tools	**Optimized application libraries**
name (type; functionality; runs on)	node libraries (type; functionality)
Performance analysis tools	parallel libraries (type; functionality)
name (type; functionality; runs on)	**Possible handicaps**
Portability tools	
name (type; functionality; runs on)	

The group *application software* is a difficult topic. One has to be careful with manufacturer lists of available software, sometimes a package is running on a single node only. Parallelized and optimized versions of existing application packages are very scarce, which - from an economic point of view - is quite understandable. And the user base for new massively parallel applications has been small until now. The group *configurations* reveals what the manufacturers offer as minimal, typical and maximal system configuration. Last but not least the group *performance* tries to provide some typical computation and communication performance data derived from available documentation.

References

[1] T. Bönniger, R. Esser, and D. Krekel, CRAY Y-MP, NEC SX-3 und Siemens VP S, Vergleichende Darstellung von Höchstleistungsrechnern, Wirtschaftsinformatik (März 1990) 273-286.

[2] T. Bönniger, R. Esser, and D. Krekel, CM-5, KSR1, and Paragon XP/S: a comparative description of massively parallel computers on the basis of a catalog of classifying characteristics, Zentralinstitut für Angewandte Mathematik, Forschungszentrum Jülich, Interner Bericht KFA-ZAM-IB-9320, to appear in Parallel Computing.

[3] T. Bönniger, R. Esser, and D. Krekel, Catalog of classifying characteristics for massively parallel computers: CM-5, KSR1, and Paragon XP/S, Arbeitsbericht RRZK-9302, Regionales Rechenzentrum an der Universität zu Köln, Dez. 1993

[4] A. Choudhary, Parallel I/O Systems, Journal of Parallel and Distributed Computing 17 (1993) 1-3.

[5] A. S. Tanenbaum, Distributed Operating Systems Anno 1992. What Have We Learned So Far? in: Distributed Computing, Practice and Experience, Proceedings of the Autumn 1992 OpenForum Technical Conference, Utrecht, November 23-27, 1992, 1-14.

This article was processed using the LaTeX macro package with LLNCS style

Appendix

Classifying characteristics

Date of survey

0 Marketing information

Company

name
year of foundation
owner
turnover last three years
number of employees

MP system division

name
starting year
turnover last three years
number of employees

Previous generation MP system

name
classification
date of first delivery
number of systems sold
. industry
. government
. universities and research institutes

Current MP system

name
date of announcement
date of first delivery
number of systems sold
. industry
. government
. universities and research institutes

1 General characteristics of the MP system

MP system

global classification
scaling range (sold systems)
MP system architecture
single-stage / multi-stage (clustered) architecture
node / cluster types
max. number of nodes / clusters
memory concept
max. memory sizes
max. filesize
globally accessible diskspace
typical access
connection with front end
external networks

Cluster

cluster architecture
nodes
intra-cluster connection
shared memory size per cluster
extended memory
I/O

Processing node

node architecture
processing elements
clock frequency
node-internal communication
node-external communication
node memory
. real / virtual (page size)
. address length
. address space
. size

Input/Output

concept
I/O bandwidth

Software

operating system
. MP system
. node
file systems
languages
programming models
parallel CASE tools

2 Processing nodes

Designation by manufacturer

Functionality

Architecture

components
clock frequency

Main processor

technology
chip type
chips per processor / processor on a chip
cache
registers and buffers
integer performance (peak)
floating point performance (peak)

Data formats

classification of data representation
length
classification of arithmetic

Arithmetic coprocessors

technology
chip type
functionality
. integer performance
. floating point performance
local memory
registers and buffers

Node memory

size
chip type, cycle time and size
latency
memory bandwidth
error handling

Node-internal communication

intra-node connections
. type
. number and width
. latency
. bandwidth

Node-external communication

communication processor
. type (number)
connections between nodes
. type (number)
. latency
.. best case (packet size)
.. worst case (packet size)
. bandwidth node-to-node
.. best case (packet size)
.. worst case (packet size)
. bandwidth (one node to all other nodes)
.. best case (packet size)
.. worst case (packet size)
communication buffer
number of independent communication operations
communication without interrupt of main processing

Hardware for performance measuring

Additional special hardware

Possible handicaps and bottlenecks

3 Special nodes

Operating system support

designation by manufacturer
functionality
architecture
processor type
size of node memory
intra-node connections
latency to each proc. node
bandwidth to each proc. node
local disks
. functionality
. size

I/O support

designation by manufacturer
functionality
architecture
processor type
intra-node connections
latency to each proc. node
bandwidth to each proc. node
latency to I/O device
bandwidth to I/O device

4 Cluster

Designation by manufacturer

Architecture

fixed (hardware) or definable by operator or user

Functionality

Nodes

number of processing nodes
nodes with special functionality

Memory

concept of cluster memory
functionality
size
chip type and size
latency
memory bandwidth
error handling

Cluster-internal communication

type
number and width
latency
bandwidth

Cluster-external communication

communication processor
. type (number)
connections between clusters
. type (number)
. latency
.. best case (packet size)
.. worst case (packet size)
. bandwidth per connection
.. best case (packet size)
.. worst case (packet size)
. bandwidth total
.. best case (packet size)
.. worst case (packet size)
communication buffer
number of independent communication operations
communication without interrupt of cluster internal processing

Additional special hardware

Fault tolerance

concept
nodes
communication

5 MP system

Nodes/clusters

number of processing nodes/clusters
nodes with spec. functionality

Memory

shared memory
. size
. chip type and size
. latency
. memory bandwidth
. error handling
extended memory
. functionality
. size
. chip type and size
. latency
. memory bandwidth
. error handling

Communication

concept
control network
. topology
. functionality
. latency (packet size)
. bandwidth
. low level communication routines
data network
. topology
. functionality
. latency (packet size)
. bandwidth
. communication techniques
. low level communication routines
diagnostic network
. topology
. functionality
. latency (packet size)
. bandwidth
. low level communication routines

Backplane

functionality
size

Typical access

internal front end
external front end

Operation requirements

power consumption
cooling (mode, type and quantity)
floor space

Additional special hardware

Fault tolerance

concept
nodes/cluster
communication

6 Front end (FE)

Designation by manufacturer

Concept

Number of FE

Type

Functionality

provide operating system for MP system
perform I/O for MP system
compilation
linking / loading
interactive access

Connection to MP system

type
bandwidth
connection to node, cluster, partition...

Compatibility with the MP system concerning data formats

7 Input/Output

I/O processor

designation by manufacturer
functionality
architecture
processor type
latency
bandwidth
control and data paths from processing node to peripherals or external networks

Disks and disk arrays

type (capacity, bandwidth)
functionality
disk striping (number of disks, bandwidth)

Archive system

type (medium, capacity, bandwidth)
functionality

Tapes

type (medium, capacity, bandwidth)

Compatibility of binary files

Possible handicaps and bottlenecks

8 Operating system (OS)

Name and type

Distributed OS

concept
MP system OS (size)
cluster OS (size)
node OS (size)
front end OS (size)

Functionality

restrictions on node level
extensions on MP system level
interactive access to FE, node, cluster ...
data center support functions
. multi-level security
. checkpoint/restart
. accounting
. disk quotas
. resource limits
. user data base
batch job scheduling
timesharing
user support functions
. virtual processors
. virtual shared memory
. user checkpoint/restart
execution scheduling
partitioning
. static or dynamic reservation by users/operator
. partition unit
. partition sizes (number of nodes)
. static or dynamic reservation of special nodes
data migration
fault tolerance
. static or dynamic error detection and replacement of a broken node or cluster
. minimal inactive partition, if a node or cluster aborts or is not bootable
. restart capability
. other error detection and recovery

File system

max. file size (length of inode, block size)
transparency of files from/to other systems

Possible handicaps and bottlenecks

9 Compilers

Fortran

standards
running on
parallelization / vectorization language constructs and directives
automatic parallelization / automatic vectorization
asynchronous I/O

C

standards
running on
parallelization / vectorization language constructs and directives
automatic parallelization / automatic vectorization
asynchronous I/O

C++

standards
running on
parallelization / vectorization language constructs and directives
automatic parallelization / automatic vectorization
asynchronous I/O

Assembler

running on
parallelization / vectorization language constructs and directives

Other compilers

name
standards
running on
parallelization / vectorization language constructs and directives
automatic parallelization / automatic vectorization

10 Parallel programming concepts, tools, and libraries

Support for programming models

SIMD parallel model
Data parallel model
MIMD shared memory model
Message passing model

Program development tools

name (type, functionality, running on)

Debugging tools

name (type; functionality; running on)

Performance analysis tools

name (type; functionality; running on)

Portability tools

name (type; functionality; running on)

Other tools

Native communication libraries

message passing
informational routines
global synchronization
global operations
remote message passing

Optimized application libraries

node libraries (type; functionality)
parallel libraries (type; functionality)

Possible handicaps

11 Application Software

DCE applications

Graphics

X Window System
generic display
realtime animation

Data base management systems

Parallelized application packages

Software catalog available

12 Configurations

Minimum configuration

number of nodes
number and types of special nodes
number of clusters
shared memory
memory per node/cluster
front end
mass storage

Typical configuration

number of nodes
number and types of special nodes
number of clusters
shared memory
memory per node/cluster
front end
mass storage

Maximum configuration (sold)

number of nodes
number and types of special nodes
number of clusters
shared memory
memory per node/cluster
front end
mass storage

13 Performance

Node performance (64-bit)

BLAS-3 (matrix multiplication 1000x1000)
LINPACK (100x100)

Computational performance of a typical configuration for linear equations (64-bit) [7]

number of processors
peak performance (r_{peak})
performance of the largest problem running on this system (r_{max})
size of the largest problem (n_{max})
size of the problem reaching the performance 1/2*r-max ($n_{1/2}$)
LINPACK (1000x1000)

Communication performance for the typical configuration between two nodes (Fortran example)

packet size
latency of a communication
. neighboring nodes
. worst case
. average
. broadcast
bandwidth of a communication
. neighbor
. worst case
. average
. broadcast

Communication performance between mass storage and node/cluster for the typical configuration

packet size
latency of a communication
. best case
. worst case
. average
bandwidth of a communication
. best case
. worst case
. average

Bisectional communication bandwidth for the typical configuration

Machine granularity

neighboring nodes
worst case
average

Computational intensity for the floating point add and floating point multiply operations [2]