Table of Contents

- Introduction and Motivation
- Theoretical Foundations
- Distributed Programming Languages
- Distributed Operating Systems
- Distributed Communication
- Distributed Data Management
- Reliability
- Applications
- Conclusions
- Appendix
Three Issues

- Use of multiple PEs
- Cooperation among the PEs
- Potential for survival to partial failure
## Control Mechanisms

<table>
<thead>
<tr>
<th>Statement type \ Control type</th>
<th>Sequential control</th>
<th>Parallel Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential/parallel statement</td>
<td><code>Begin S₁, S₂ end</code></td>
<td><code>Parbegin S₁, S₂, Parend, Fork/join</code></td>
</tr>
<tr>
<td>Alternative statement</td>
<td><code>goto, case if C then S₁ else S₂</code></td>
<td>Guarded commands: $G \rightarrow C$</td>
</tr>
<tr>
<td>Repetitive statement</td>
<td><code>for … do</code></td>
<td><code>doall, for all</code></td>
</tr>
<tr>
<td>Subprogram</td>
<td><code>procedure Subroutine</code></td>
<td><code>procedure subroutine</code></td>
</tr>
</tbody>
</table>

Four basic sequential control mechanisms with their parallel counterparts.
Focus 6: Expressing Parallelism

parbegin/parend statement

\[ S_1;[[S_2;[S_3||S_4];S_5;S_6]]|S_7];S_8 \]

A precedence graph of eight statements.
**Focus 6 (Cont’d.)**

**fork/join statement**

\[
\begin{align*}
s_1; \\
c_1 & := 2; \\
\text{fork } L1; \\
s_2; \\
c_2 & := 2; \\
\text{fork } L2; \\
s_4; \\
\text{go to } L3; \\
L1: & s_3; \\
L2: & \text{join } c_1; \\
& s_5; \\
L3: & \text{join } c_2; \\
& s_6;
\end{align*}
\]

A precedence graph.
Dijkstra's Semaphore + Parbegin/Parend

$S(i)$: A sequence of $P$ operations; $S_i$; a sequence of $V$ operations

$s$: a binary semaphore initialized to 0.

$S(1): S_1; V(s_{12}); V(s_{13})$
$S(2): P(s_{12}); S_2; V(s_{24}); V(s_{25})$
$S(3): P(s_{13}); S_3; V(s_{35})$
$S(4): P(s_{24}); S_4; V(s_{46})$
$S(5): P(s_{25}); P(s_{35}); S_5; V(s_{56})$
$S(6): P(s_{46}); P(s_{56}); S_6$
Focus 7: Concurrent Execution

- $R(S_i)$, the **read set** for $S_i$, is the set of all variables whose values are referenced in $S_i$.
- $W(S_i)$, the **write set** for $S_i$, is the set of all variables whose values are changed in $S_i$.
- **Bernstein conditions:**
  - $R(S_1) \cap W(S_2) = \varnothing$
  - $W(S_1) \cap R(S_2) = \varnothing$
  - $W(S_1) \cap W(S_2) = \varnothing$
Example 7

\[ S_1 : a := x + y, \]
\[ S_2 : b := x \times z, \]
\[ S_3 : c := y - 1, \text{ and} \]
\[ S_4 : x := y + z. \]

\[ S_1 \parallel S_2, S_1 \parallel S_3, S_2 \parallel S_3, \text{ and } S_3 \parallel S_4. \]

Then, \( S_1 \parallel S_2 \parallel S_3 \) forms a largest complete subgraph.
Example 7 (Cont’d.)

A graph model for Bernstein's conditions.
Alternative Statement

Alternative statement in DCDL (CSP like distributed control description language)

\[ [ G_1 \rightarrow C_1 \square G_2 \rightarrow C_2 \square \ldots \square G_n \rightarrow C_n ] . \]
Example 8

Calculate $m = \max \{x, y\}$:

$\begin{align*}
[x \geq y \rightarrow m := x \quad \Box \quad y \geq x \rightarrow m := y]\end{align*}$
Repetitive Statement

\[*[ G_1 \rightarrow C_1 \square G_2 \rightarrow C_2 \square \ldots \square G_n \rightarrow C_n ].*\]
Example 9

meeting-time-scheduling ::= t := 0;

*[ t := a(t) □ t := b(t) □ t := c(t) ]
Communication and Synchronization

- One-way communication: send and receive
- Two-way communication: RPC(Sun), RMI(Java and CORBA), and rendezvous (Ada)

Several design decisions:
- One-to one or one-to-many
- Synchronous or asynchronous
- One-way or two-way communication
- Direct or indirect communication
- Automatic or explicit buffering
- Implicit or explicit receiving
- Persistent (message-queueing) or transient communication
- Discrete or streaming communication (syn. and QoS)
<table>
<thead>
<tr>
<th>Primitives</th>
<th>Example Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PARALLELISM</strong></td>
<td></td>
</tr>
<tr>
<td>Expressing parallelism</td>
<td></td>
</tr>
<tr>
<td>Processes</td>
<td>Ada, Concurrent C, Lina, NIL, Emerald, Concurrent Smalltalk</td>
</tr>
<tr>
<td>Objects</td>
<td>Occam</td>
</tr>
<tr>
<td>Statements</td>
<td>Par Alfl, FX-87</td>
</tr>
<tr>
<td><strong>Expressions</strong></td>
<td>Concurrent PROLOG, PARLOG</td>
</tr>
<tr>
<td>Clauses</td>
<td></td>
</tr>
<tr>
<td><strong>Mapping</strong></td>
<td></td>
</tr>
<tr>
<td>Static</td>
<td>Occam, Star Mod</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Concurrent PROLOG, ParAlfl</td>
</tr>
<tr>
<td>Migration</td>
<td>Emerald</td>
</tr>
<tr>
<td><strong>COMMUNICATION</strong></td>
<td></td>
</tr>
<tr>
<td>Message Passing</td>
<td></td>
</tr>
<tr>
<td>Point-to-point messages</td>
<td>CSP, Occam, NIL</td>
</tr>
<tr>
<td>Rendezvous</td>
<td>Ada, Concurrent C</td>
</tr>
<tr>
<td>Remote procedure call</td>
<td>DP, Concurrent CLU, LYNX</td>
</tr>
<tr>
<td>One-to-many messages</td>
<td>BSP, StarMod</td>
</tr>
<tr>
<td>Data Sharing</td>
<td></td>
</tr>
<tr>
<td>Distributed data Structures</td>
<td>Lina, Orca</td>
</tr>
<tr>
<td>Shared logical variables</td>
<td>Concurrent PROLOG, PARLOG</td>
</tr>
<tr>
<td><strong>Nondeterminism</strong></td>
<td></td>
</tr>
<tr>
<td>Select statement</td>
<td>CSP, Occam, Ada, Concurrent C, SR</td>
</tr>
<tr>
<td>Guarded Horn clauses</td>
<td>Concurrent PROLOG, PARLOG</td>
</tr>
<tr>
<td><strong>PARTIAL FILURES</strong></td>
<td></td>
</tr>
<tr>
<td>Failure detection</td>
<td>Ada, SR</td>
</tr>
<tr>
<td>Atomic transactions</td>
<td>Argus, Aeolus, Avalon</td>
</tr>
<tr>
<td>NIL</td>
<td></td>
</tr>
</tbody>
</table>
Message-Passing Library for Cluster Machines (e.g., Beowulf clusters)

- Parallel Virtual Machine (PVM):
  [www.epm.ornl/pvm/pvm_home.html](http://www.epm.ornl/pvm/pvm_home.html)

- Message Passing Interface (MPI):
  [www.mpi.nd.edu/lam/](http://www.mpi.nd.edu/lam/)

- Java multithread programming:
  [www.mcs.drexel.edu/~shartley/ConcProjJava](http://www.mcs.drexel.edu/~shartley/ConcProjJava)
  [www.ora.com/catalog/jenut](http://www.ora.com/catalog/jenut)

- Beowulf clusters:
  [www.beowulf.org](http://www.beowulf.org)
Message-Passing (Cont’d.)

- **Asynchronous** point-to-point message passing:
  - send message list to destination
  - receive message list \{from source\}

- **Synchronous** point-to-point message passing:
  - send message list to destination
  - receive empty signal from destination
  - receive message list from sender
  - send empty signal to sender
Example 10

The squash program replaces every pair of consecutive asterisks "**" by an upward arrow “↑”.

input::= * [ send c to squash ]
output::= * [ receive c from squash ]
Example 10 (Cont’d.)

squash::= 
  *[ receive c from input →  
      [ c ≠ * → send c to output  
        □ [ c = * → receive c from input;  
          [ c ≠ * → send * to output;  
            send c to output  
            c = * → send ↑ to output  
          ] □  
        ] ]  
    ] ]
Focus 8: Fibonacci Numbers

- \( F(i) = F(i-1) + F(i - 2) \) for \( i > 1 \), with initial values \( F(0) = 0 \) and \( F(1) = 1 \).

- \( F(i) = (\phi^i - \phi'^i)/ (\phi - \phi') \), where \( \phi = (1 + 5^{0.5})/2 \) (golden ratio) and \( \phi' = (1 - 5^{0.5})/2 \).

0, 1, 2, 3, 5, 8, 13, 21, 35, 54, …
Focus 8 (Cont’d.)

A solution for $F(n)$. 
**Focus 8 (Cont’d.)**

- $f(0) ::=$
  - `send n to f(1);`
  - `receive p from f(2);`
  - `receive q from f(1);`
  - `ans := q`

- $f(-1) ::=$
  - `receive p from f(1)`
Focus 8 (Cont’d.)

- $f(i) ::= \begin{align*}
    &\text{receive } n \text{ from } f(i - 1); \\
    &\left[ \begin{array}{l}
    n > 1 \rightarrow [ \text{send } n - 1 \text{ to } f(i + 1); \\
    \quad \text{receive } p \text{ from } f(i + 2); \\
    \quad \text{receive } q \text{ from } f(i + 1); \\
    \quad \text{send } p + q \text{ to } f(i - 1); \\
    \quad \text{send } p + q \text{ to } f(i - 2) \\
    \end{array} \right] \\
\end{align*}

\begin{align*}
\square n = 1 &\rightarrow [ \text{send } 1 \text{ to } f(i - 1); \\
                  &\quad \text{send } 1 \text{ to } f(i - 2) ] \\
\square n = 0 &\rightarrow [ \text{send } 0 \text{ to } f(i - 1); \\
                  &\quad \text{send } 0 \text{ to } f(i - 2) ] \\
\end{align*}
Focus 8 (Cont’d.)

Another solution for \( F(n) \).
Focus 8 (Cont’d.)

- \( f(0) ::= \)
  
  \[
  [ \ n > 1 \rightarrow [ \text{send } n \text{ to } f(1); \\
  \text{receive } p \text{ from } f(1); \\
  \text{receive } q \text{ from } f(1); \\
  \text{ans} := p \\
  ] \\
  ]
  \]

- \( n = 1 \rightarrow \text{ans} := 1 \)
- \( n = 0 \rightarrow \text{ans} := 0 \)

]
f(i)::=

receive \( n \) from \( f(i - 1) \);

[ \( n > 1 \) → [ send \( n - 1 \) to \( f(i + 1) \);

receive \( p \) from \( f(i + 1) \);

receive \( q \) from \( f(i + 1) \);

send \( p + q \) to \( f(i - 1) \);

send \( p \) to \( f(i - 1) \)

]

\( n = 1 \) → [ send 1 to \( f(i - 1) \);

send 0 to \( f(i - 1) \)

]

]
Focus 9: Message-Passing Primitives of MPI

- MPI_Isend: asynchronous communication
- MPI_send: receipt-based synchronous communication
- MPI_ssend: delivery-based synchronous communication
- MPI_sendrecv: response-based synchronous communication
Message-passing primitives of MPI: Isend, send, ssend, sendrecv.
Focus 10: Interprocess Communication in UNIX

- Socket: int socket (int domain, int type, int protocol).
  - domain: normally internet.
  - type: datagram or stream.
  - protocol: TCP (Transport Control Protocol) or UDP (User Datagram Protocol)
- Socket address: an Internet address and a local port number.
Sockets used for datagrams

Sender

\[
\begin{align*}
& s = \text{socket(AF_INET, SOCK_DGRAM, 0)} \\
& \ldots \\
& \text{bind}(s, \text{ClientSocketAddress}) \\
& \ldots \\
& \text{sendto}(s, \text{"message"}, \text{ServerSocketAddress})
\end{align*}
\]

Receiver

\[
\begin{align*}
& t = \text{socket(AF_INET, SOCK_DGRAM, 0)} \\
& \ldots \\
& \text{bind}(t, \text{ServerSocketAddress}) \\
& \ldots \\
& \text{amount} = \text{recvfrom}(t, \text{buffer, from})
\end{align*}
\]
High-Level (Middleware) Communication Services

- Achieve access transparency in distributed systems
  - Remote procedure call (RPC)
  - Remote method invocation (RMI)
Remote Procedure Call (RPC)

- Allow programs to call procedures located on other machines.
- Traditional (synchronous) RPC and asynchronous RPC.
Remove Method Invocation (RMI)
Robustness

- Exception handling in high level languages (Ada and PL/1)
- Four Types of Communication Faults
  - A message transmitted from a node does not reach its intended destinations
  - Messages are not received in the same order as they were sent
  - A message gets corrupted during its transmission
  - A message gets replicated during its transmission
Failures in RPC

If a **remote procedure call** terminates abnormally (the time out expires) there are four possibilities.

- The receiver did not receive the call message.
- The reply message did not reach the sender.
- The receiver crashed during the call execution and either has remained crashed or is not resuming the execution after crash recovery.
- The receiver is still executing the call, in which case the execution could interfere with subsequent activities of the client.
Exercise 3

1.(The Welfare Crook by W. Feijen) Suppose we have three long magnetic tapes each containing a list of names in alphabetical order. The first list contains the names of people working at IBM Yorktown, the second the names of students at Columbia University and the third the names of all people on welfare in New York City. All three lists are endless so no upper bounds are given. It is known that at least one person is on all three lists. Write a program to locate the first such person (the one with the alphabetically smallest name). Your solution should use three processes, one for each tape.
Exercise 3 (Cont’d.)

2. Convert the following DCDL expression to a precedence graph.

\[
[ S_1 \parallel [ [ S_2 \parallel S_3 ]; S_4 ] \parallel S_5 ]
\]

Use **fork** and **join** to express this expression.

3. Convert the following program to a precedence graph:

\[
S_1;[[S_2;S_3\parallel S_4;S_5\parallel S_6]\parallel S_7];S_8
\]
Exercise 3 (Cont’d.)

4. $G$ is a sequence of integers defined by the recurrence $G_i = G_{i-1} + G_{i-3}$ for $i > 1$, with initial values $G_0 = 0$, $G_1 = 1$, and $G_2 = 1$. Provide a DCDL implementation of $G_i$ and use one process for each $G_i$.

5. Using DCDL to write a program that replaces $a*b$ by $a \uparrow b$ and $a**b$ by $a \downarrow b$, where $a$ and $b$ are any characters other than * . For example, if $a_1a_2*a_3**a_4***a_5$ is the input string then $a_1a_2 \uparrow a_3 \downarrow a_4***a_5$ will be the output string.