Compilación Optimizante para el Lenguaje Occam sobre un Sistema Time-Sharing

Guillermo A. Alvarez * Daniel O. Bandinelli *

Marcelo O. Fernández * Gustavo A. Rosini *

Resumen: Occam es un lenguaje de programación basado en un modelo de procesos secuenciales que se ejecutan concurrentemente, comunicándose por medio de pasaje sincrónico de mensajes. Se detallan los aspectos del desarrollo de un compilador Occam–C sobre una máquina uniprocesador. El sistema operativo UNIX es usado como plataforma de implementación. Se incluyen explicaciones detalladas de los chequeos realizados, de los procedimientos de optimización, y de las técnicas de compilación desarrolladas. Se define la traducción de un lenguaje diseñado para arquitecturas multiprocesador sobre sistemas monoprocesador time-sharing.
Palabras clave: Compiladores, Generación Optimizante de Código, Lenguajes de Programación, Programación Concurrente, Sistemas Operativos.

1 Introducción

Durante las últimas dos décadas, diferentes paradigmas de programación paralela han sido propuestos. Entre ellos se encuentran los paradigmas basados en pasaje de mensajes como Occam [7]. Esta familia de lenguajes tiene su sustento teórico en CSP [5], y admite como soporte a arquitecturas multiprocesador con memoria distribuida y pasaje de mensajes.

Este trabajo describe el desarrollo de un compilador para el lenguaje concurrente Occam que genera código C sobre el sistema operativo UNIX System V. Si bien el lenguaje Occam fue concebido originalmente para ser ejecutado en un array de procesadores homogéneos (Transputers), en nuestro caso se simula la ejecución paralela de procesos secuenciales utilizando las facilidades provistas por UNIX para programación concurrente y comunicación entre procesos.

El compilador acepta un subconjunto del lenguaje Occam estándar con la adición de extensiones destinadas a facilitar la interacción con el usuario. Dichas extensiones permiten entrada/salida respetando el modelo conceptual del lenguaje.

La compilación de un programa consiste en dos etapas que se realizan secuencialmente. La primera (*front-end*) acepta como entrada el texto de un programa Occam sintácticamente correcto que cumple las restricciones inherentes al lenguaje, y lo transforma en código para una máquina abstracta. La segunda (*back-end*) toma la salida de la primera y genera código para una máquina real en términos del código para la máquina abstracta. Cada una de estas etapas se efectúa en una pasada sobre el código, y por lo tanto el proceso total de compilación insume solamente dos pasadas.

El trabajo está organizado como sigue. Las Secciones 2 y 3 detallan aspectos del front-end y del back-end, respectivamente. En la Sección 4 se explican las optimizaciones implementadas; por último, en la Sección 5 se presentan las conclusiones.

*Escuela Superior Latino Americana de Informática (ESLAI) – CC 3193 – (1000) Buenos Aires – Argentina; e-mail alvarezg@buevm2.vnet.ibm.com, danielb@dcfcen.edu.ar

2 Chequeos y Generación de Código Intermedio

En esta sección se detallarán los diversos chequeos realizados sobre el programa Occam de entrada. Para realizar dichos chequeos, se debe recoger durante la fase de análisis del texto del programa la siguiente información relativa a los identificadores: nombre, tipo (variable, canal o constante), valor (para las constantes), alcance, e información de posición. La información de posición denota qué tipo de accesos al identificador se hacen y en qué lugares (procesos) del texto fuente se los hace. Es usada primordialmente para garantizar que la comunicación entre procesos se realize utilizando canales y no variables compartidas. Además sirve para chequear la correcta utilización de los canales.

Por intermedio de los chequeos realizados hemos podido determinar tanto errores fatales como warnings que ayudan al programador a realizar un código más confiable. Los chequeos realizados son detallados a continuación.

Utilización de variables compartidas: No permitir que dos procesos que pueden ejecutarse en paralelo se comuniquen a través de variables compartidas.

Utilización de canales: Dados dos procesos que pueden ejecutarse en paralelo, verificar que sólo uno envía y sólo uno recibe mensajes por un canal.

Constantes, variables y canales no utilizados: Se detectan los identificadores declarados pero no utilizados en el texto Occam, notificándose con el warning correspondiente.

Variable que es leída y no está inicializada: También se avisa (a través de un warning) de esta situación para prevenir al programador de un comportamiento posiblemente no determinístico del programa.

Los identificadores utilizados deben estar declarados una sola vez: Se chequea que todo identificador esté declarado en el alcance en que se lo utiliza. Además no permitimos declaraciones de un identificador cuando ya haya sido declarado en el mismo alcance con un tipo distinto. La violación de cualquiera de estos dos requisitos resulta en un error.

Uso de un identificador de acuerdo a su tipo: Se verifica que los identificadores se usen consistentemente con sus declaraciones, por ejemplo que en una escritura por canal no se pretenda escribir el dato sobre una variable entera.

La expresión que aparece en un proceso `WAIT` debe ser una comparación de tiempos: Sólo se permiten procesos `WAIT` de la forma `WAIT NOW AFTER exp` donde `exp` no contiene ocurrencias de `NOW`.

Un solo tipo de operador asociativo debe aparecer en una expresión no parentizada: No se permite que una secuencia de expresiones esté operada con operadores asociativos distintos, a menos que esté máximamente parentizada.

Para escribir el front-end se utilizaron el generador de analizadores lexicográficos `lex` y el generador de parsers `yacc` provistos por UNIX. Estos brindan una forma clara y cercana al modelo teórico de especificar el lenguaje que se desea reconocer usando esquemas de traducción [1]. Además permiten modificar fácilmente la gramática, lo cual no ocurriría si tanto el parser como el analizador lexicográfico estuvieran implementados directamente en C. El hecho de utilizarlos para el desarrollo del compilador favorece su portabilidad a otros sistemas UNIX.

El código intermedio generado como consecuencia de la primera etapa de la compilación tiene la forma de un árbol sintáctico, que refleja la estructura del programa fuente después de que se le realizan las optimizaciones detalladas en la Sección 4.

3 Generación de Código

La presente sección contiene aspectos relativos a las prestaciones que brinda el soporte de tiempo de ejecución implementado como complemento del compilador, y a la generación de código C a partir de los principales constructores Occam.

3.1 Soporte de Tiempo de Ejecución

La traducción de Occam a C involucra la utilización de facilidades de sincronización y comunicación entre procesos, por lo cual el soporte de tiempo de ejecución permite crear y liberar dinámicamente dichos recursos en forma eficiente. Esto se logra a través de manejadores que se encargan de administrar el uso de canales de comunicación, variables en memoria compartida y semáforos.

El manejador de canales se construyó usando las facilidades provistas por los manejadores de semáforos y de memoria compartida. La comunicación entre procesos es unidireccional, es decir hay un proceso emisor del mensaje (en Occam este mensaje es un entero) y el otro proceso es el receptor, no pudiéndose intercambiar ambos papeles. Además la comunicación es sincrónica, esto implica que se lleva a cabo cuando ambos procesos (el emisor y el receptor) quieren comunicarse. Si sólo uno de los procesos está listo para la comunicación debe esperar a que el otro también lo esté.

De los identificadores Occam, se marcan como residentes en memoria compartida sólo a aquellos que probablemente se usen para comunicar resultados de un proceso hijo a su padre. Esto ocurre cuando un componente de un constructor `PAR` modifica una variable que es leída posteriormente por un proceso que se ejecuta después de que el `PAR` termina. Dado que UNIX no permite crear variables compartidas individuales (porque se debe crear un segmento de memoria compartida cuyo tamaño debe ser mayor que un límite fijado por el sistema operativo [3]), se debe implementar un manejador intermedio que permita efectuar pedidos de memoria con mayor frecuencia y granularidad más fina. Como las variables compartidas van a ser usadas por varios procesos simultáneamente, las estructuras utilizadas por el manejador deben estar en memoria compartida para poder ser accedidas por todos los procesos.

Para manejar semáforos se han definido operaciones según la notación usual (definida en [4]). Se debe tener en cuenta que UNIX [9] no maneja semáforos individuales sino arreglos de los mismos. Este problema es similar al visto anteriormente para administrar las variables compartidas, y se soluciona implementando un manejador que sigue los mismos principios.

Las variables compartidas y los canales del programa Occam determinan la generación de dos clases de código: declaraciones para los identificadores C asociados a ellos, y código ejecutable C que invoca a los respectivos manejadores, creando y liberando los recursos cuando los correspondientes identificadores entran y salen de alcance.

3.2 Estructura del Generador de Código

Dado un árbol que representa el código intermedio del programa Occam, se lo recorre recursivamente generando el correspondiente código C. Para ello se genera un encabezamiento que produce la inclusión de los manejadores de semáforos, canales y variables compartidas, y además declara algunas variables globales del soporte de tiempo de ejecución. Luego se genera la declaración de la función `main()` donde se inicializan los manejadores y las variables del

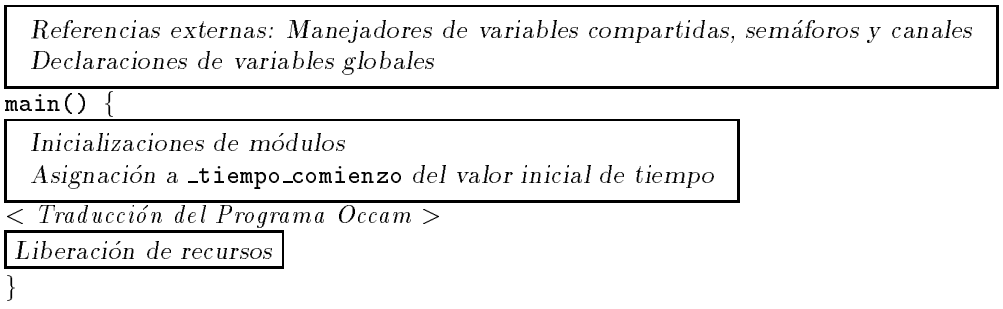

Figura 1: Formato Típico del Código C Generado.

soporte, y a continuación el código correspondiente al árbol; se termina con el cierre de los módulos utilizados.

El formato de un programa C típico, generado como salida del back-end, es mostrado en la Figura 1.

Declaraciones: Al procesar el código intermedio correspondiente a un bloque de declaraciones Occam, se genera un bloque C donde las variables Occam se declaran como variables enteras si no son compartidas. Para los canales, generamos una declaración de tipo `_chan`.

Después de generar las declaraciones del bloque C y antes de comenzar con el código, se generan los `create_vc()` y los `create_chan()` correspondientes, para que los manejadores creen los recursos en tiempo de ejecución. Análogamente, al terminar el procesamiento del bloque, generamos los `remove_vc()` y `remove_chan()` correspondientes antes de cerrar el bloque C.

No es necesario generar declaraciones para las constantes Occam en esta etapa, ya que al construir el árbol las apariciones de las constantes se reemplazaron por el valor correspondiente.

Cabe notar que se generan únicamente las declaraciones correspondientes a aquellas variables y canales que realmente fueron usados, y no para los que fueron declarados pero no utilizados, lográndose de esta manera el consecuente ahorro de espacio de almacenamiento y de costo de creación al ejecutar.

Constructor PAR: La traducción del constructor `PAR` consiste en generar código que crea dinámicamente procesos UNIX, cada uno de los cuales ejecuta uno de los procesos componentes del constructor. La creación de procesos se hace en forma de árbol binario completo por niveles para mejorar la eficiencia, como se explicará en la Sección 4.

Por lo tanto, una construcción Occam de la forma `PAR` P_1 ... P_m se traducirá, si $m > 1$ (si no, se genera código directamente para el único proceso), en el código C de la Figura 2.

La variable `_n` indica al generador la cantidad de procesos a crear; a la salida, `_i` contiene el número de orden de cada proceso generado, por lo cual puede elegir la rama correcta del `switch()` y ejecutar su código. Cada proceso genera a lo sumo dos hijos (según su posición en el árbol), ejecuta su rama del `PAR` y luego espera que sus hijos terminen. El proceso padre ejecuta la primera componente del constructor. Dentro del código del generador se tienen en cuenta posibles errores de ejecución al excederse la cantidad máxima de procesos permitida por el sistema.

Constructor ALT: Si alguna de las guardas del `ALT` se revela verdadera en tiempo de compilación, entonces se genera solamente el código del proceso guardado correspondiente, como una implementación del no-determinismo.

La política que se adoptará al evaluar las guardas será considerarlas como una generalización

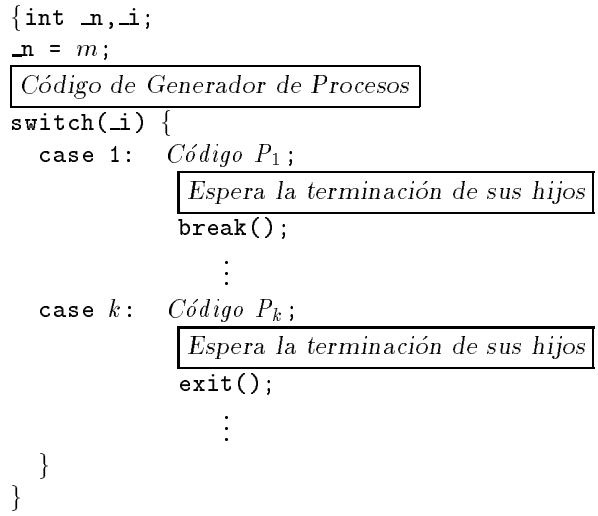

Figura 2: Traducción del constructor `PAR`.

del concepto de expresiones booleanas: una guarda (en el caso más complejo) será cierta cuando la expresión booleana que está a la izquierda del `&` sea cierta, y además la operación que está a la derecha pueda realizarse. En este contexto, esta expresión se evaluará siguiendo la semántica estándar de Occam, de izquierda a derecha y con corto circuito. Por lo tanto, al ejecutarse el código generado a partir de un `ALT`, se evaluarán primero todas las expresiones que figuren en las partes izquierdas de guardas, y se descartarán definitivamente las que den `FALSE`. Esta decisión posibilita una implementación en la que ninguno de los casos posibles provoca espera activa, como se verá más adelante, ya que aún en el caso de expresiones que involucren a la función primitiva `NOW` su valor de verdad se evaluará una sola vez.

Tras este paso inicial de evaluación, pueden existir guardas de la forma `exp & SKIP` cuya expresión haya evaluado a `TRUE`, y que por lo tanto sean ciertas. Es por esto que se genera código para verificar al ejecutar si esto ocurre, y en el caso de que ocurra se dice que el `ALT` se ha *trivializado*: se ejecutará el proceso guardado correspondiente.

Las guardas que pasaron la primera etapa (sus expresiones son verdaderas, o bien consisten en una entrada por canal o `WAIT` aislados) son de dos clases distintas, asumiendo que el `ALT` no se ha trivializado: entradas o procesos `WAIT`. De aquí en adelante sólo nos referiremos a estas guardas. Para implementar su evaluación, se genera en tiempo de corrida un proceso para cada una de estas guardas. La generación de procesos se hace en forma plana y no de árbol binario ya que si se mantuviera la estructura de creación de árbol binario, podría pasar que algunos procesos fueran descartados por ser falsas las expresiones de sus guardas, pero éstos no generarían a sus hijos en el orden arbóreo, que sí podrían tener guardas válidas. Dicha generación de procesos se hace sólo en los casos en que el `ALT` no se ha trivializado y no todas las guardas tienen expresiones falsas como componente.

Ninguno de estos procesos hace espera activa, ya que los encargados de evaluar `WAIT`s calculan la cantidad de tiempo que tendrá que pasar y ejecutan un `sleep()` de esa duración; por otra parte, los encargados de evaluar entradas se duermen mientras no haya un dato disponible en el canal. Una implementación alternativa para el `ALT` hubiera sido evaluar continuamente las guardas hasta que alguna se cumpliera, desperdiciando tiempo de CPU en actividades inútiles. Fue necesario definir un mecanismo para que el proceso cuya guarda se cumpla primero avise al

proceso padre (coordinador) que debe ejecutar su correspondiente proceso guardado. Para ello, el padre genera primero los procesos evaluadores de guardas y luego se queda suspendido en el semáforo `_alguien_termino` (que inicialmente vale 0). Cuando un proceso hijo puede cumplir su guarda, verifica primero si puede avisarle al padre, para garantizar que el padre aceptará a sólo un proceso; si puede, dejará su número en un espacio de memoria compartida, y luego ejecutará un `V(_alguien_termino)` despertando por lo tanto al padre. El padre, una vez enterado de que todos sus hijos instalaron sus manejadores (a través del semáforo `_activados`), espera hasta que uno de ellos triunfe; sólo entonces interrumpe enviando señales a los demás procesos hijos para que éstos terminen su ejecución. Esto se hace para no esperar a guardas demasiado lentas, y para reducir la cantidad de procesos en el sistema al mínimo. Un hijo que satisface su guarda ejecuta un `P(_primero_que_termina)` (inicialmente en 1), para que si dos guardas se cumplen al mismo tiempo sólo una pueda avisar al padre y la otra quede suspendida en dicho semáforo.

Los procesos que esperan entradas sobre el mismo canal necesitan que sólo uno de ellos retire un dato de ese canal; para lograr esto, se define un semáforo auxiliar por cada canal (`_sem_`*nomcanal*) inicialmente en 1, que sólo permite que una guarda ejecute la lectura sobre dicho canal.

Los procesos que esperan entradas sobre canales distintos deben mantener la semántica de Occam en el sentido de que sólo el que logra ejecutar su proceso guardado retira el dato del canal, dejando los demás canales sin cambio (en ellos no se retira el dato ni se desbloquea al proceso que lo envió), y además tampoco altera las variables que figuran como destino en las entradas no exitosas. Para solucionar el problema, se define un `receive_guard()` sobre canales, el cual no desbloquea al proceso que envió el dato hasta haberle avisado al padre que él es la guarda verdadera. Y para no alterar variables destino de entradas supuestamente no ejecutadas, se definen variables auxiliares sobre las cuales realmente se hace la entrada. Al comienzo del proceso guardado, se asignan estas variables a la verdadera variable destino. Por lo tanto, la evaluación de guardas a cargo de procesos independientes no causa efectos laterales indeseables.

Cuando un proceso evaluador de guarda es interrumpido por su padre, pasa a ejecutar un manejador, que en el caso de los `WAIT` y de las entradas desde el `stdin` se limita a abortar la ejecución, pero en el caso de las entradas desde canales Occam se fija si estaba a la mitad de la lectura, restaurando en ese caso el estado del canal.

Por lo tanto, una construcción Occam de la forma `ALT` $\begin{matrix} G_1 & & G_m \\ & \ldots & \\ P_1 & & P_m \end{matrix}$ se traducirá en el código C de la Figura 3, donde *nomvar* es la variable sobre la cual se realiza la entrada.

Demás Constructores: Los constructores `SEQ`, `IF` y `WHILE` se traducen de la forma natural. En cuanto a las lecturas y escrituras de canales, se realizan a través de las primitivas sincrónicas del manejador de canales, salvo en el caso de `IN` y `OUT`. Para éstos se generan llamadas a las rutinas estándares de C para entrada/salida.

4 Optimizaciones

En esta sección se mencionan las optimizaciones implementadas como parte del compilador. Dichas optimizaciones se dividen básicamente entre las que se realizan en tiempo de compilación (específicamente dentro del front-end) y las que se realizan en tiempo de ejecución (consecuencia del código generado especialmente por el back-end).

Como optimizaciones implementadas en la primera etapa pueden destacarse las siguientes:

```
{int _i,_n,_proc_guard = 0,_alguna_exp_cierta = 0,_alt_simplificado = 0;
_semaforo _alguien_termino,_primero_que_termina,_activados;
_vc _termino;
Declaraciones de Otras Variables y Semáforos Auxiliares
create_vc(&_termino);
create_sem(&_alguien_termino,0);
create_sem(&_primero_que_termina,1);
Creación e Inicialización de Variables y Semáforos Auxiliares
_n = m ;
Evaluación de las Expresiones Booleanas
if (!_alt_simplificado || _alguna_exp_cierta){
   create_sem(&_activados,m);
   Generador Plano de Procesos
   switch(_i){
       case 1: Código G1
               exit();
                  ⋮
       case m + 1:P(&_alguien_termino);
                  _proc_guard = read_vc(_termino);
                  wait_for_zero(&_activados);
                  Enviar Interrupción a los Hijos no Exitosos
                  Esperar Fin de Hijos
       }
   }
switch(_proc_guard) {
   case 1: Código P1
           break;
            ⋮
   case m: Código Pm
           break;
   }
remove_vc(_termino);
Liberación de Variables y Semáforos
}
```

(a)

```
signal(ABORT,_handler_wait);
P(&_activados);
Código de WAIT
P(&_primero_que_termina);
write_vc(_termino,_i);
V(&_alguien_termino);
```

(b)

```
signal(ABORT,_handler_entrada);
P(&_activados);
P(&_sem_nomcanal);
receive_guard(&_nomcanal,&_primero_que_termina,
  &_alguien_termina,_nomvar,_i);
```

(c)

Figura 3: Constructor `ALT`: (a) Código Principal; (b) Código de un Proceso Evaluador de Guardas `WAIT`; (c) Código de un Proceso Evaluador de Guardas con Entradas desde Canales.

- Evaluación y propagación estática de constantes, teniendo en cuenta los aspectos de evaluación de izquierda a derecha (corto circuito) y de coerción de entero a booleano.
- Simplificación del código fuente, eliminando estructuras de control cuyo comportamiento es predecible estáticamente (por ejemplo, un `WHILE` cuya condición evalúa estáticamente a `FALSE`) y manejando los casos de construcciones vacías. Se manejan los casos en que la eliminación de partes inútiles de código convierte a su vez en inútiles a los constructores que las encierran. Esto se hace en base a la evaluación estática de constantes.
- Simplificación de estructuras anidadas con constructores del mismo tipo, para evitar el uso excesivo de recursos atribuibles a la implementación de dichos constructores.
- Minimización del uso de variables en memoria compartida, marcando como tales sólo a las que probablemente se usen para comunicar resultados de un proceso hijo a su padre. Esto ocurre cuando un componente de un constructor `PAR` modifica una variable que es leída por un proceso que se ejecuta después de que el `PAR` termina. Se descartó la alternativa de mantener todas las variables en memoria compartida porque implicaba un nivel adicional de indirección para cada acceso, más las llamadas al manejador de memoria compartida para reservar y liberar espacio.
- Minimización del costo por llamadas al manejador de memoria dinámica de C durante la compilación, implementando un nivel adicional que sigue una heurística orientada a no liberar memoria hasta que sea altamente probable que no se la necesite de nuevo. Esto es utilizado para hacer más eficiente el procesamiento de los bloques de declaraciones.
- Construcciones que extienden a Occam, mediante canales de comunicación con la entrada y la salida estándares de UNIX.
- Generación de código intermedio declarando identificadores sólo para aquellos que efectivamente se usan en el código Occam.
- Mensajes de error y warning destinados a facilitar el proceso de depuración de programas Occam y el uso racional de recursos, avisando de la eventual no terminación de programas, variables no utilizadas, variables no inicializadas, etc.

en tanto que en la segunda etapa se destacan:

- Utilización de manejadores para disminuir la cantidad de llamadas al sistema operativo aumentando la eficiencia, y también para proveer servicios de la granularidad necesaria para la ejecución del código C generado.
- Generación paralela asincrónica de procesos en tiempo de corrida, utilizando un esquema recursivo que extrae el máximo paralelismo del sistema time-sharing. Básicamente, en vez de hacer que el proceso padre genere a todos sus hijos y espere a que terminen para seguir, cada proceso genera a 0, 1 o 2 hijos y luego realiza su parte del cómputo. Debido a que la creación de un proceso involucra entrada/salida de periféricos del sistema, esta estrategia mejora la eficiencia creando los procesos según una estructura de árbol binario que aprovecha el paralelismo entre cómputo y entrada/salida.

- Minimización de la cantidad de procesos generados dinámicamente, de dos formas distintas: haciendo que el proceso padre en un constructor `PAR` ejecute una parte del código en vez de esperar a que sus hijos terminen, y evaluando sólo las guardas del `ALT` que tengan alguna posibilidad de hacerse ciertas.
- Minimización de la cantidad de procesos existentes simultáneamente en el sistema y de los tiempos de espera para sincronizar, a través de la emisión de interrupciones del proceso controlador del `ALT` a los evaluadores de guardas que los obliga a terminar tan pronto como alguna guarda se ha hecho cierta.

5 Conclusiones

Se desarrolló un compilador que hace dos pasadas, la primera sobre el código fuente Occam generando directamente un árbol correcto, chequeado y optimizado; y la segunda sobre el código intermedio generado por la primera. Se implementaron optimizaciones que influyen tanto en la eficiencia de ejecución del programa Occam como en el uso de recursos provistos por el sistema operativo. Como extensiones a Occam, se definieron canales que comunican a un proceso con la entrada y la salida estándares de UNIX.

Referencias

[1] A. Aho, R. Sethi, J. Ullman. *Compilers: Principles, Techniques and Tools*. Addison-Wesley, 1986.

[2] P. Ancilotti, M. Boari. *Programmazione Concorrente*. Versione Preliminare, 1987.

[3] M. Bach. *The Design of the UNIX Operating System*. Prentice-Hall Int., 1986.

[4] E. W. Dijkstra. Cooperating Sequential Processes. Technical Report EWD-123, Technological University, Eindhoven, The Netherlands, 1965.

[5] C. Hoare. *Communicating Sequential Processes*. Prentice-Hall Int., 1985.

[6] M. E. C. Hull. Occam - A Programming Language for Multiprocessor Systems. *Computer Languages*, Volume 12, Number 1, pp. 27-37.

[7] INMOS Limited. *OCCAM Programming Manual*. Prentice-Hall Int., 1984.

[8] J. Peterson, A. Silberschatz. *Operating Systems Concepts*. Addison-Wesley, 1985.

[9] W. R. Stevens. *UNIX Network Programming*. Prentice-Hall Int., 1990.