UNIVERSIDAD DE BUENOS AIRES

TESIS DE GRADO DE INGENIERÍA EN INFORMÁTICA

Detección de Deadlocks en Rust en tiempo de compilación mediante Redes de Petri

Autor: Horacio Lisdero Scaffino
hlisdero@fi.uba.ar

Director: Ing. Pablo Andrés Deymonnaz
pdeymon@fi.uba.ar

Departamento de Computación
Facultad de Ingeniería

8 de junio de 2023

Índice general

Índice de figuras

List of Listings

Siglas

ART Android Runtime

AST abstract syntax tree

BB basic blocks

CFG control flow graph

CLI command-line interface

CPN Colored Petri nets

CPU central processing unit

Creol Concurrent Reflective Object-oriented Language

CTL* Computational Tree Logic*

DBMS Database management systems

FSM Finite-state machine

HIR High-Level Intermediate Representation

IR intermediate representation

ISA instruction set architecture

JIT just-in-time

LHS left-hand side

LIFO last in, first out

LoLA Low-Level Petri Net Analyzer

LTO link time optimization

MCC Model Checking Contest

MIR Mid-level Intermediate Representation

NT-PN Nondeterministic Transitioning Petri nets

OOM out-of-memory

OS operating system

P/T nets place/transition nets

PIPE2 Platform Independent Petri net Editor 2

PN Petri nets

PNML Petri Net Markup Language

RAG Resource Allocation Graph

RAII Resource Acquisition Is Initialization

RFCs Requests for Comments

RHS right-hand side

TAPAAL Tool for Verification of Timed-Arc Petri Nets

THIR Typed High-Level Intermediate Representation

TWF transaction-wait-for

UB Undefined Behavior

WASM WebAssembly

XML Extensible Markup Language

Abstract

Detección de Deadlocks en Rust en tiempo de compilación mediante Redes de Petri

En la presente tesis de grado se presenta una herramienta de análisis estático para detección de *deadlocks* y señales perdidas en el lenguaje de programación Rust. Se realiza una traducción en tiempo de compilación del código fuente a una red de Petri. Se obtiene entonces la red de Petri como salida en uno o más de los siguientes formatos: DOT, Petri Net Markup Language o LoLA. Posteriormente se utiliza el verificador de modelos LoLA para probar de forma exhaustiva la ausencia de *deadlocks* y de señales perdidas. La herramienta está publicada como *plugin* para el gestor de paquetes *cargo* y la totalidad del código fuente se encuentra disponible en GitHub[1][2]. La herramienta demuestra de forma práctica la posibilidad de extender el compilador de Rust con un pase adicional para detectar más clases de errores en tiempo de compilación.

Compile-time Deadlock Detection in Rust using Petri Nets

This undergraduate thesis presents a static analysis tool for the detection of deadlocks and missed signals in the Rust programming language. A compile-time translation of the source code into a Petri net is performed. The Petri net is then obtained as output in one or more of the following formats: DOT, Petri Net Markup Language, or LoLA. Subsequently, the LoLA model checker is used to exhaustively prove the absence of deadlocks and missed signals. The tool is published as a plugin for the package manager *cargo* and the entirety of the source code is available on GitHub[1][2]. The tool demonstrates in a practical way the possibility to extend the Rust compiler with an additional pass to detect more error classes at compile time.

[1] https://github.com/hlisdero/cargo-check-deadlock/
[2] https://github.com/hlisdero/netcrab

Capítulo 1

Introducción

Para comprender plenamente el alcance y el contexto de este trabajo, es beneficioso proporcionar algunos temas de fondo que sientan las bases de la investigación. Estos temas de fondo sirven como bloques teóricos sobre los que se construye la traducción.

En primer lugar, se presenta la teoría de las redes de Petri tanto gráficamente como en términos matemáticos. Para ilustrar el poder de modelado y la versatilidad de las redes de Petri, se proporcionan al lector varios modelos diferentes a modo de ejemplo. Estos modelos muestran la capacidad de las redes de Petri para capturar diversos aspectos de los sistemas concurrentes y representarlos de forma visual e intuitiva. Más adelante, se introducen algunas propiedades importantes y se explica el análisis de alcanzabilidad que realiza el verificador de modelos.

En segundo lugar, se analiza brevemente el lenguaje de programación Rust y sus principales características. Se incluye un puñado de ejemplos de aplicaciones notables de Rust en la industria. Se reúnen pruebas convincentes del uso de lenguajes con un manejo seguro de la memoria para argumentar que Rust proporciona una base excelente para ampliar la detección de clases de errores en tiempo de compilación.

En tercer lugar, se ofrece información general sobre el problema de los bloqueos mutuos y las señales perdidas cuando se utilizan *condition variables*, así como una descripción de las estrategias habituales utilizadas para resolver estos problemas.

Por último, se ofrece una visión general de la arquitectura de los compiladores y del concepto de verificación de modelos. Señalaremos el potencial aún sin explorar que subyace a la verificación formal para aumentar la seguridad y fiabilidad de los sistemas de software.

1.1. Redes de Petri

1.1.1. Visión general

Las redes de Petri (Petri nets (PN)) son una herramienta de modelado gráfico y matemático utilizada para describir y analizar el comportamiento de los sistemas concurrentes. Fueron introducidas por el investigador alemán Carl Adam Petri en su tesis doctoral [Petri, 1962] y desde entonces se han aplicado en diversos campos como la informática, la ingeniería y la biología. Puede encontrar un resumen conciso de la teoría de las redes de Petri, sus propiedades, análisis y aplicaciones en [Murata, 1989].

Una red de Petri es un grafo dirigido bipartito formado por un conjunto de lugares, transiciones y arcos. Hay dos tipos de nodos: lugares y transiciones. Los lugares representan el estado del sistema, mientras que las transiciones representan eventos o acciones que pueden ocurrir. Los arcos conectan lugares a transiciones o transiciones a lugares. No puede haber arcos entre dos lugares o entre dos transiciones, preservando así la propiedad bipartita.

Los lugares pueden contener cero o más marcas o fichas. Los tokens se utilizan para representar la presencia o ausencia de entidades en el sistema, como recursos, datos o procesos. En la clase más simple de redes de Petri, los tokens no llevan ninguna información y son indistinguibles unos de otros. El número de fichas en un lugar o la simple presencia de una ficha es lo que transmite significado en la red. Las fichas se consumen y se producen al dispararse las transiciones, lo que da la impresión de que se mueven a través de los arcos.

En la representación gráfica convencional, los lugares se representan mediante círculos, mientras que las transiciones se representan como rectángulos. Las fichas se representan como puntos negros dentro de los lugares, como se ve en la Fig. 1.1.

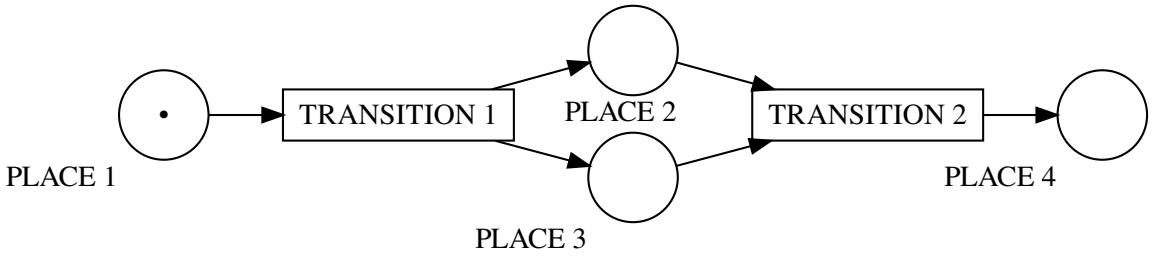

Figura 1.1: Ejemplo de una red de Petri. PLACE 1 contiene una marca.

Cuando una transición se dispara, consume fichas de sus lugares de entrada y produce fichas en sus lugares de salida, lo que refleja un cambio en el estado del sistema. El disparo de una transición se activa cuando hay suficientes fichas en sus lugares de entrada. En la Fig. 1.2, podemos ver cómo se producen los disparos uno detrás del otro.

El disparo de las transiciones habilitadas no es determinista, es decir, se disparan aleatoriamente

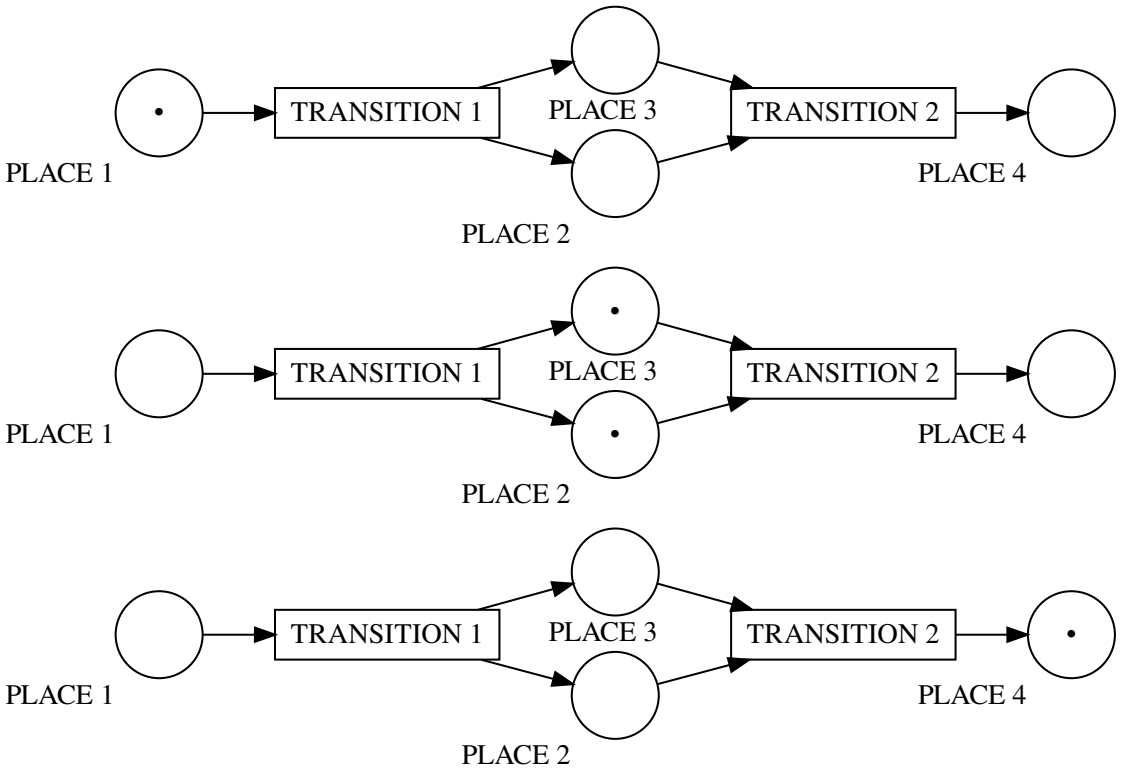

Figura 1.2: Ejemplo de disparo de una transición: La transición 1 se dispara primero, luego se dispara la transición 2.

mientras estén habilitadas. Una transición deshabilitada se considera *muerta* si no hay ningún estado alcanzable en el sistema que pueda llevar a que la transición se habilite. Si todas las transiciones de la red están muertas, entonces la red también se considera *muerta*. Este estado es análogo al bloqueo (*deadlock*) de un programa informático.

Las redes de Petri pueden utilizarse para modelar y analizar una amplia gama de sistemas, desde sistemas sencillos con unos pocos componentes hasta sistemas complejos con muchos componentes que interactúan entre sí. Pueden utilizarse para detectar problemas posibles en un sistema, optimizar su rendimiento y diseñar e implementar sistemas de forma más eficaz.

También pueden utilizarse para modelar procesos industriales [Van der Aalst, 1994], para validar requisitos de software expresados como casos de uso [Silva and Dos Santos, 2004] o para especificar y analizar sistemas en tiempo real [Kavi et al., 1996].

En concreto, las redes de Petri pueden utilizarse para detectar deadlocks en el código fuente modelando el programa de entrada como una red de Petri y analizando después la estructura de la red resultante. Se demostrará que este enfoque es formalmente sólido y practicable para el código fuente escrito en el lenguaje de programación Rust.

1.1.2. Modelo matemático formal

Una red de Petri es un tipo particular de grafo bipartito, con pesos y dirigido, dotado de un estado inicial denominado *marcado inicial*, M_0. Para este trabajo, se utilizará la siguiente definición general de una red de Petri tomada de [Murata, 1989].

Definition 1: Petri net

Una red de Petri es una 5-tupla, $PN = (P, T, F, W, M_0)$ *donde:*
$P = \{p_1, p_2, \ldots, p_m\}$ *es un conjunto finito de lugares,*
$T = \{t_1, t_2, \ldots, t_n\}$ *es un conjunto finito de transiciones,*
$F \subseteq (P \times T) \cup (T \times P)$ *es un conjunto de arcos (relación de flujo),*
$W : F \leftarrow \{1, 2, 3, ...\}$ *es una función de peso para los arcos,*
$M_0 : P \leftarrow \{0, 1, 2, 3, ...\}$ *es el marcado inicial,*
$P \cap T = \varnothing$ *y* $P \cup T \neq \varnothing$

En la representación gráfica, los arcos se etiquetan con su peso que es un número entero no negativo k. Normalmente el peso se omite si es igual a 1. Un arco con peso k puede interpretarse como un conjunto de k arcos paralelos distintos.

Un *marcado (estado)* asocia a cada lugar un número entero no negativo l. Si un marcado asigna al lugar p un número entero no negativo l, decimos que p está *marcado con* l *marcas o tokens*. Pictóricamente, denotamos esto colocando l puntos negros (fichas) en el lugar p. El *p-ésimo* componente de M, denotado por $M(p)$, es el número de fichas en el lugar p.

Una definición alternativa de las redes de Petri utiliza un multiconjunto (*bag*) en lugar de un conjunto para definir los arcos, permitiendo así la presencia de múltiples elementos. Puede encontrarse en la literatura, por ejemplo, [Peterson, 1981, Definition 2.3].

Como ejemplo, consideremos la red de Petri $PN_1 = (P, T, F, W, M)$ donde:

$P = \{p_1, p_2\}$,
$T = \{t_1, t_2\}$,
$F = \{(p_1, t_1), (p_2, t_2), (t_1, p_2), (t_2, p_2)\}$,
$W(a_i) = 1 \quad \forall a_i \in F$
$M(p_1) = 0, M(p_2) = 0$

Esta red no contiene fichas y todos los pesos de los arcos son iguales a 1. Se muestra en la Fig. 1.3.

La Fig. 1.3 contiene una estructura interesante que encontraremos más adelante. Esto motiva la siguiente definición.

Figura 1.3: Example of a small Petri net containing a self-loop.

Definition 2: Bucle

Un lugar p y una transición t definen un bucle si p es a la vez un lugar de entrada y un lugar de salida de t.

En la mayoría de los casos, nos interesan las redes de Petri que no contienen bucles las cuales se denominan *puras*.

Definition 3: Red de Petri pura

Se dice que una red de Petri es pura si no tiene bucles.

Además, si el peso de cada arco es igual a uno, llamamos a la red de Petri *ordinaria*.

Definition 4: Red de Petri ordinaria

Se dice que una red de Petri es ordinaria si todos los pesos de sus arcos son 1, es decir,

$$W(a) = 1 \quad \forall a \in F$$

1.1.3. Disparo de transiciones

La regla de disparo de transición es el concepto central de las redes de Petri. A pesar de ser aparentemente simple, sus implicaciones son de gran alcance y complejidad.

Definition 5: Regla de disparo de transiciones

Sea $PN = (P, T, F, W, M_0)$ una red de Petri.

(I) Se dice que una transición t está habilitada si cada lugar de entrada p de t marcado con al menos $W(p,t)$ marcas donde $W(p,t)$ es el peso del arco que va de p de t.

(II) Una transición activada puede dispararse o no, dependiendo de si el evento tiene lugar o no.

(III) El disparo de una transición activada t elimina $W(t,p)$ marcas de cada lugar de entrada p de t donde $W(t,p)$ es el peso del arco de t a p.

Siempre que se habiliten varias transiciones para un marcado M dado, puede dispararse cualquiera de ellas. La elección es no determinista. Se dice que dos transiciones habilitadas están en *conflicto* si el disparo de una de ellas inhabilita la otra transición. En este caso, las transiciones compiten por la ficha colocada en un lugar de entrada compartido.

Si dos transiciones t_1 y t_2 están habilitadas en algún marcado pero no están en conflicto, pueden dispararse en cualquier orden, es decir, t_1 luego t_2 o t_2 luego t_1. Tales transiciones representan eventos que pueden ocurrir concurrentemente o en paralelo. En este sentido, el modelo de red de Petri adopta un modelo de paralelismo basado en el intercalado (*interleaved model of parallelism*), es decir, el comportamiento del sistema es el resultado de un intercalado arbitrario de los eventos paralelos.

Las transiciones sin lugares de entrada ni lugares de salida reciben un nombre especial.

Definition 6: Transición fuente (*Source transition*)

Una transición sin ningún lugar de entrada se denomina transición fuente.

Definition 7: Transición sumidero (*Sink transition*)

Una transición sin lugar de salida se denomina transición de sumidero.

Cabe destacar que una transición fuente se activa incondicionalmente y produce fichas sin consumir ninguna, mientras que el disparo de una transición sumidero consume fichas sin producir ninguna.

1.1.4. Simuladores en línea

Para familiarizarse con la dinámica de las redes de Petri, resulta útil simular algunos ejemplos en línea, ya que ver una red de Petri en acción es más claro que cualquier explicación estática sobre el papel. Hemos reunido algunas herramientas con este fin para aliviar la carga del lector.

- Puede encontrar un sencillo simulador hecho por Igor Kim en https://petri.hp102.ru/. La herramienta incluye un vídeo tutorial en Youtube y redes de ejemplo.
- Como complemento, el profesor Wil van der Aalst de la Universidad de Hamburgo ha elaborado una serie de tutoriales interactivos. Estos tutoriales son archivos de Adobe Flash Player (con extensión `.swf`) que los navegadores web modernos no pueden ejecutar. Por suerte, se puede utilizar un emulador Flash en línea como el que se encuentra en https://flashplayer.fullstacks.net/?kind=Flash_Emulator para cargar los archivos y ejecutarlos.
- Otro editor y simulador de redes de Petri en línea es http://www.biregal.com/. El usuario puede dibujar la red, añadir los tokens y luego disparar manualmente las transiciones.

1.1.5. Ejemplos de modelado

En esta subsección, se presentan varios ejemplos sencillos para introducir algunos conceptos básicos de las redes de Petri que son útiles en el modelado. Esta subsección se ha adaptado de [Murata, 1989].

Para otros ejemplos de modelado, como el problema de exclusión mutua, los semáforos propuestos por Edsger W. Dijkstra, el problema del productor/consumidor y el problema de los filósofos cenando, se remite al lector a [Peterson, 1981, Chap. 3] y [Reisig, 2013].

Máquinas de estado finito

Las máquinas de estados finitos (Finite-state machine (FSM)) pueden representarse mediante una subclase de redes de Petri.

Como ejemplo de máquina de estado finito, consideremos una máquina expendedora de café. Acepta monedas de 1 € o 2 € euros y vende dos tipos de café, el primero cuesta 3 € euros y el segundo 4 € euros. Supongamos que la máquina puede contener hasta 4 € y no devuelve ningún cambio. Entonces, el diagrama de estados de la máquina puede representarse mediante la red de Petri que se muestra en la Fig. 1.4.

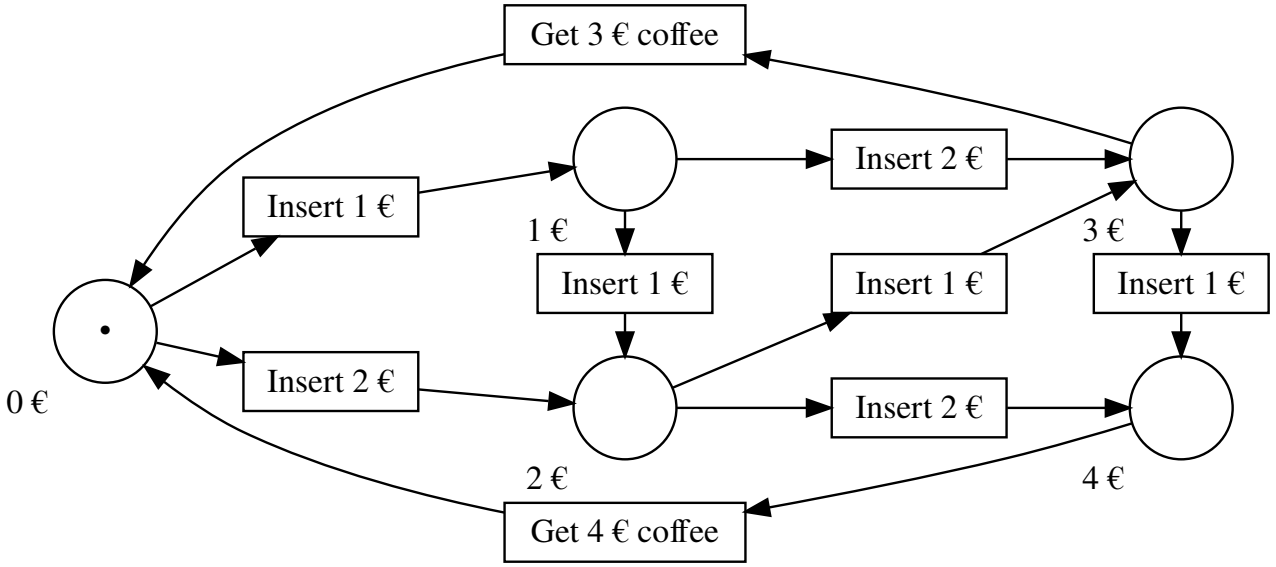

Figura 1.4: La red de Petri para una máquina expendedora de café, es equivalente a un diagrama de estados.

Las transiciones representan la inserción de una moneda del valor etiquetado, por ejemplo, "Insert 1 € coin". Los lugares representan un posible estado de la máquina, es decir, la cantidad de dinero almacenada actualmente en su interior. El lugar situado más a la izquierda, etiquetado "0 €", está marcado con una ficha y corresponde al estado inicial del sistema.

Ahora podemos presentar la siguiente definición de esta subclase de redes de Petri.

Definition 8: Máquinas de estado

Una red de Petri en la que cada transición tiene exactamente un arco entrante y exactamente un arco saliente se conoce como máquina de estados.
Cualquier FSM (o su diagrama de estados) puede modelarse con una máquina de estados.

La estructura de un lugar p_1 que tiene dos (o más) transiciones de salida t_1 y t_2 se denomina conflicto, decisión o elección, según la aplicación en cuestión. Esto se ve en el lugar inicial de la Fig. 1.4, donde el usuario debe seleccionar qué moneda introducir al principio.

Actividades en paralelo

A diferencia de las máquinas de estados finitos, las redes de Petri también pueden modelar actividades paralelas o concurrentes. En la Fig. 1.5 se muestra un ejemplo de ello, en el que la red representa la división de una tarea mayor en dos subtareas que pueden ejecutarse en paralelo.

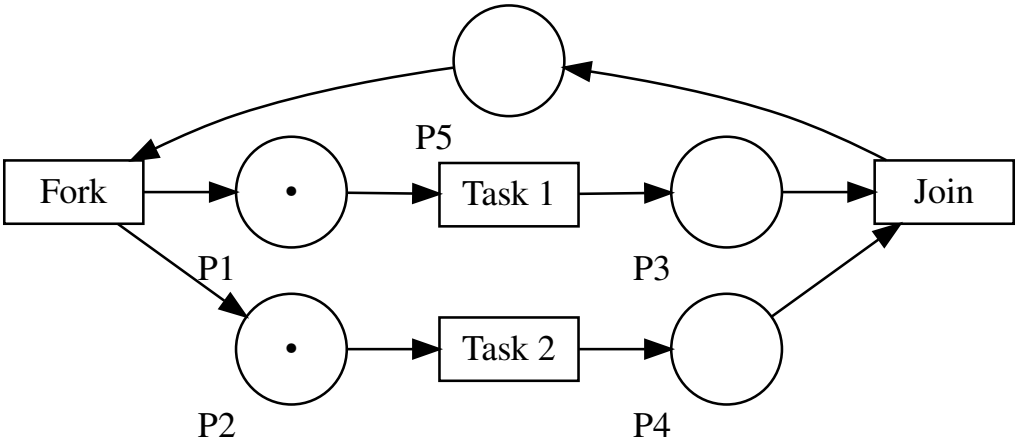

Figura 1.5: La red de Petri que representa dos actividades paralelas en forma de bifurcación.

La transición "Fork" se disparará antes que "Task 1" y "Task 2" y que "Join" sólo se disparará después de ambas tareas se completan. Pero tenga en cuenta que el orden en que se ejecutan la "Task 1" y la "Task 1" no es determinista. La tarea 1 podría dispararse antes, después o al mismo tiempo que la tarea 2. Es precisamente esta propiedad de la regla de disparo en las redes de Petri la que permite modelar sistemas concurrentes.

Definition 9: Concurrencia en redes de Petri

Se dice que dos transiciones son concurrentes si son causalmente independientes, es decir, el disparo de una transición no causa ni es provocado por el disparo de la otra.

Observe que cada lugar de la red de la Fig. 1.5 tiene exactamente un arco entrante y un arco saliente. Esta subclase de redes de Petri permite representar la concurrencia pero no las decisiones (conflictos).

Definition 10: Grafos marcados (*marked graphs*)

Una red de Petri en la que cada lugar tiene exactamente un arco entrante y exactamente un arco saliente se conoce como grafo marcado (marked graph).

Protocolos de comunicación

Los protocolos de comunicación también pueden representarse en redes de Petri. La fig. 1.6 ilustra un protocolo sencillo en el que el Proceso 1 envía mensajes al Proceso 2 y espera a recibir un acuse de recibo antes de continuar. Ambos procesos se comunican a través de un canal con búfer cuya capacidad máxima es de un mensaje. Por lo tanto, sólo un mensaje puede estar viajando entre los procesos en un momento dado. Para simplificar, no se ha incluido ningún mecanismo de *timeout*.

Se podría incorporar al modelo un tiempo de espera máximo para la operación de envío añadiendo una transición $t_{timeout}$ con aristas de "Wait for ACK" a "Ready to send". Esto mapea la decisión entre recibir el acuse de recibo y el tiempo de espera.

Control de sincronización

En un sistema multihilo, los recursos y la información se comparten entre varios hilos. Esta compartición debe controlarse o sincronizarse para garantizar el correcto funcionamiento del sistema global. Las redes de Petri se han utilizado para modelar diversos mecanismos de sincronización, incluidos los problemas de exclusión mutua, lectores-escritores y productores-consumidores [Murata, 1989].

En la Fig. 1.7 se muestra una red de Petri para un sistema de lectores-escritores con k procesos. Cada marca representa un proceso y la elección de T1 o T2 representa si el proceso realiza una operación de lectura o de escritura.

Utiliza aristas ponderadas para eliminar atómicamente $k-1$ tokens de P3 antes de realizar una escritura (transición T2), evitando así que los lectores entren en el bucle derecho de la red.

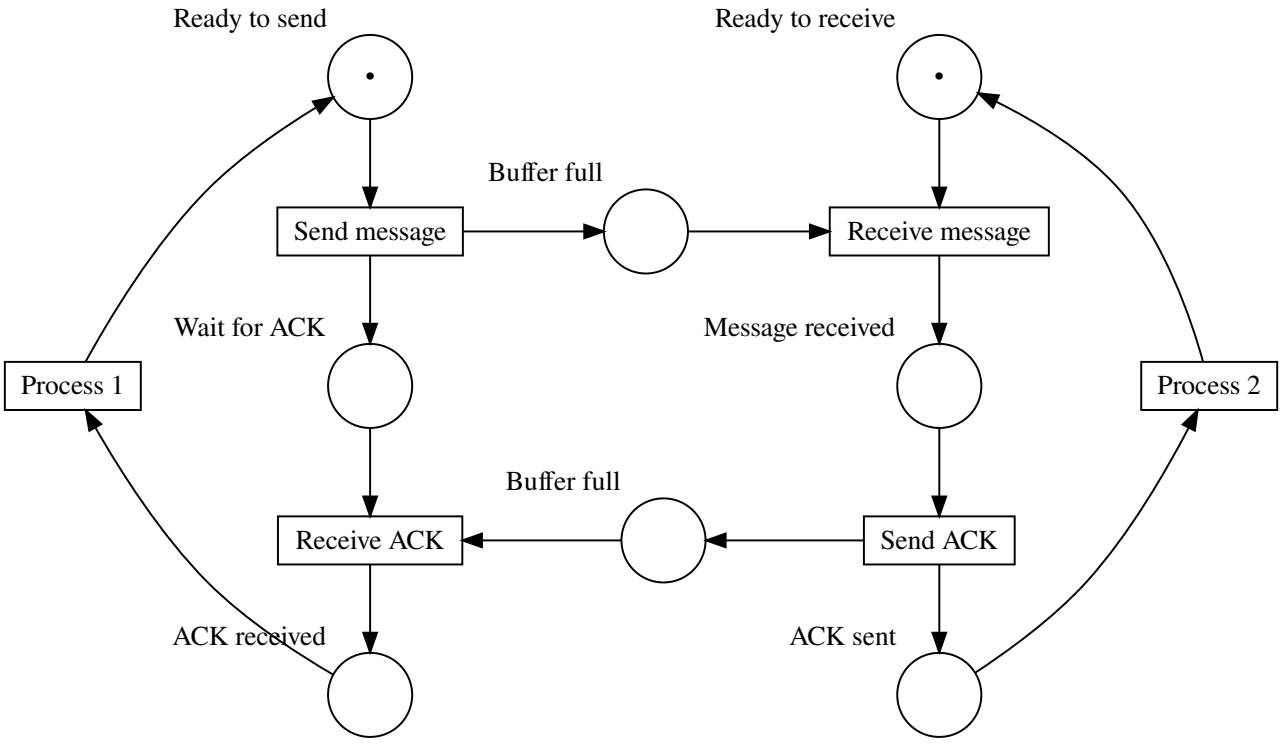

Figura 1.6: Modelo simplificado de red de Petri de un protocolo de comunicación.

Como máximo k procesos pueden estar leyendo al mismo tiempo, pero cuando un proceso esté leyendo, ningún proceso podrá leer, es decir, P2 estará vacío. Se puede comprobar fácilmente que la propiedad de exclusión mutua se satisface para el sistema.

Hay que señalar que este sistema no está libre de inanición (*starvation*), ya que no hay garantía de que una operación de escritura vaya a producirse en algún momento. Por otro lado, el sistema sí está libre de deadlocks.

1.1.6. Propiedades importantes

En esta subsección veremos conceptos fundamentales para el análisis de redes de Petri que facilitarán la comprensión de las redes que trataremos en el resto del trabajo.

Alcanzabilidad

La alcanzabilidad es una de las cuestiones más importantes cuando se estudian las propiedades dinámicas de un sistema. El disparo de transiciones habilitadas provoca cambios en la ubicación de las marcas. En otras palabras, cambia el marcado M. Una secuencia de disparos crea una

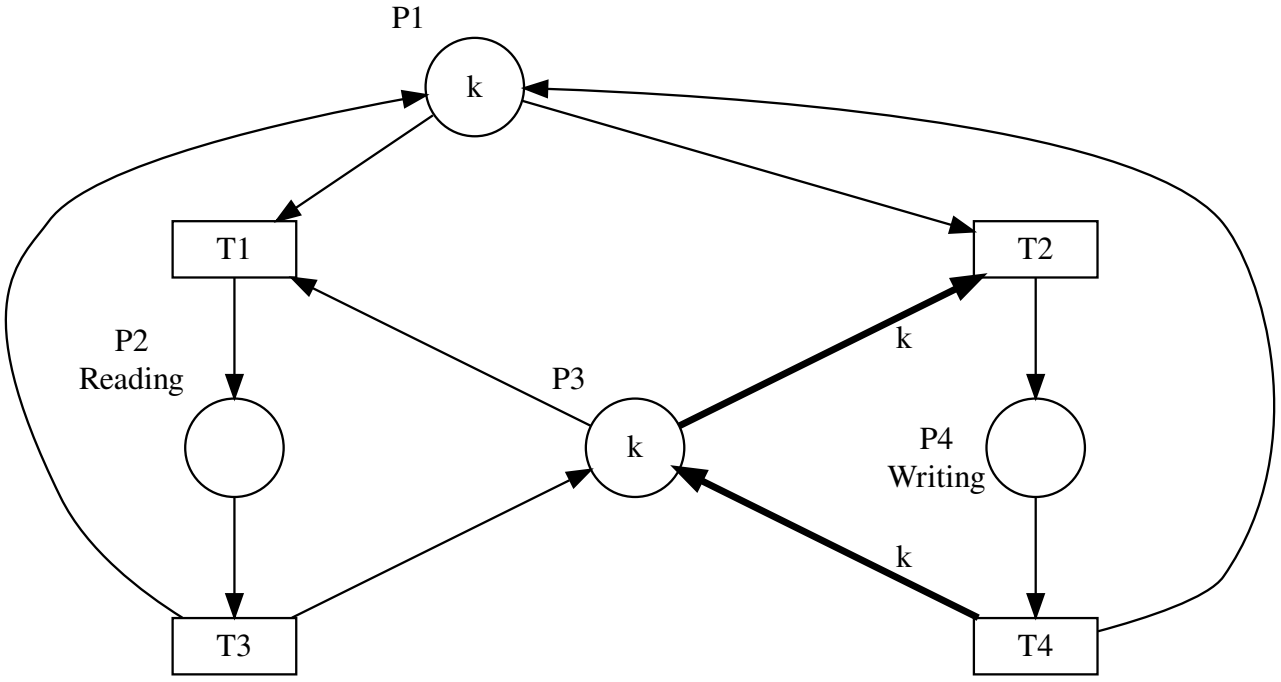

Figura 1.7: Un sistema de redes de Petri con k procesos que leen o escriben.

secuencia de marcados en la que cada marcado puede denotarse como un vector de longitud n, siendo n el número de lugares de la red de Petri.

Una *secuencia de disparos* u *ocurrencias* se denota por $\sigma = M_0\ t_1\ M_1\ t_2\ M_2\ \cdots\ t_l\ M_l$ o simplemente $\sigma = t_1\ t_2\ \cdots\ t_l$, ya que las marcas resultantes de cada disparo se derivan de la regla de disparo de transición descrita en la Sec. 1.1.3.

Definition 11: Alcanzabilidad (*Reachability*)

Decimos que una marca M es alcanzable desde M_0 si existe una secuencia de disparo σ tal que M está contenida en σ.

El conjunto de todas las marcas posibles alcanzables desde M_0 se denota por $R(N, M_0)$ o más sencillamente $R(M_0)$ cuando la red en cuestión es obvia. Este conjunto se denomina *conjunto de alcanzabilidad*.

Se puede presentar entonces un problema de suma importancia en la teoría de las redes de Petri, a saber, el *Problema de alcanzabilidad*: Encontrar si $M_n \in R(M_0, N)$ para una red y un marcado inicial dados.

En algunas aplicaciones, sólo nos interesan los marcados de un subconjunto de lugares y podemos ignorar los restantes. Esto da lugar a una variación del problema conocida como *problema de alcanzabilidad del submarcado* (*submarking reachability problem*).

Se ha demostrado que el problema de la alcanzabilidad es decidible [Mayr, 1981]. Sin embargo, también se ha demostrado que ocupa un espacio exponencial (formalmente, es EXPSPACE-hard) [Lipton, 1976]. Se han propuesto nuevos métodos para que los algoritmos sean más eficientes [Küngas, 2005]. Recientemente, [Czerwiński et al., 2020] mejoraron el límite inferior y demostraron que el problema no es ELEMENTARY. Estos resultados ponen de relieve que el problema de la alcanzabilidad sigue siendo un área activa de investigación en teória de la computación.

Para éste y otros problemas clave, los resultados teóricos más importantes obtenidos hasta 1998 se detallan en [Esparza and Nielsen, 1994].

Acotamiento y seguridad

Durante la ejecución de una red de Petri, los tokens pueden acumularse en algunos lugares. Diversas aplicaciones sulen necesitar garantizar que el número de fichas en un lugar determinado no supere una cierta tolerancia. Por ejemplo, si un lugar representa un búfer, nos interesa que el búfer nunca se desborde.

Definition 12: Acotamiento (*Boundedness*)

Un lugar de una red de Petri es k-acotada o es k-seguro si el número de fichas de ese lugar no puede superar un número entero finito k para cualquier marcado alcanzable desde M_0.
Una red de Petri es k-acotada o simplemente acotada si todos los lugares están acotados.

La seguridad es un caso especial de la acotación. Aplica cuando el lugar contiene 1 ó 0 fichas durante la ejecución.

Definition 13: Seguridad (*Safeness*)

Un lugar de una red de Petri es seguro si el número de fichas de ese lugar nunca es superior a uno.
Una red de Petri es segura si cada lugar de esa red es seguro.

Las redes de las Fig. 1.4, 1.5 y 1.6 son todas seguras.

La red de la Fig. 1.7 es k-acotada porque todos sus lugares son k-acotados.

Liveness

El concepto de liveness es análogo a la ausencia total de deadlocks en los programas informáticos.

Definition 14: Liveness

Se dice que una red Patri (N, M_0) está viva (o equivalentemente se dice que M_0 es una marca viva (live) para N) si, para cada marca alcanzable desde M_0, es posible disparar cualquier transición de la red progresando a través de alguna secuencia de disparo.

Cuando una red está viva, siempre puede seguir ejecutándose, sin importar las transiciones que se dispararon antes. Eventualmente, cada transición puede dispararse de nuevo. Si una transición sólo puede dispararse una vez y no hay forma de volver a activarla, entonces la red no es viva (*live*).

Esto equivale a decir que la red de Petri está *libre de bloqueo* (*deadlock-free*). Definamos ahora lo que constituye un bloqueo/deadlock y mostremos ejemplos de ello.

Definition 15: Bloqueo en redes de Petri

Un bloqueo o deadlock en una red Patri es una transición (o un conjunto de transiciones) que no puede dispararse para ninguna marca alcanzable desde M_0. La transición (o un conjunto de transiciones) no puede volver a activarse después de un cierto punto de la ejecución.

Una transición está *viva* si no está bloqueada. Si una transición está viva, siempre es posible elegir una serie de disparos de transiciones adecuada para pasar del marcado actual a un marcado que habilite la transición.

Las redes de las Fig. 1.4, 1.5 and 1.6 están todas vivas. En todos estos casos, después de algunos disparos, la red vuelve al estado inicial y puede reiniciar el ciclo.

La red de la Fig. 1.1 no está viva. Después de dos disparos termina de ejecutarse y no puede ocurrir nada más. La red de la Fig. 1.3 tampoco está viva, porque T1 sólo se ejecutará una vez y a partir de ese momento sólo se podrá activar T2.

1.1.7. Análisis de alcanzabilidad

Tras haber introducido el conjunto de alcanzabilidad $R(N, M_0)$ en la sección 1.1.6, ahora podemos presentar una técnica de análisis importante para las redes de Petri: *el árbol de alcanzabilidad* (*reachability tree*).

Ejecutaremos paso a paso el algoritmo para construir el árbol de alcanzabilidad y, a continuación, presentaremos sus ventajas e inconvenientes. En términos generales, el árbol de alcanzabilidad tiene la siguiente estructura: Los nodos representan las marcas generadas a partir de M_0, la raíz del árbol y sus sucesores. Cada arco representa un disparo de transición que transforma un marcado en otro.

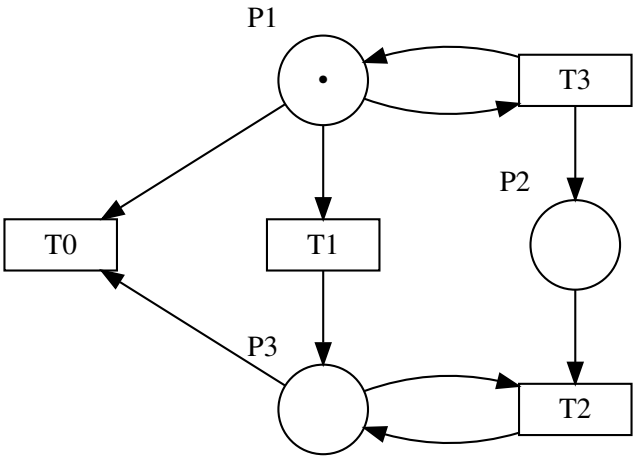

Figura 1.8: Una red de Petri marcada para ilustrar la construcción de un árbol de alcanzabilidad.

Considere la red de Petri mostrada en la Fig. 1.8. La marca inicial es $(1, 0, 0)$. En este marcado inicial se habilitan dos transiciones: T1 y T3. Dado que queremos obtener todo el conjunto de alcanzabilidad, definimos un nuevo nodo en el árbol de alcanzabilidad para cada marcado alcanzable, que resulta de disparar cada transición. Un arco, etiquetado por la transición disparada, conduce desde la marca inicial (la raíz del árbol) hasta cada una de las nuevas marcas. Tras este primer paso (Fig. 1.9), el árbol contiene todas las marcas que son inmediatamente alcanzables desde la marca inicial.

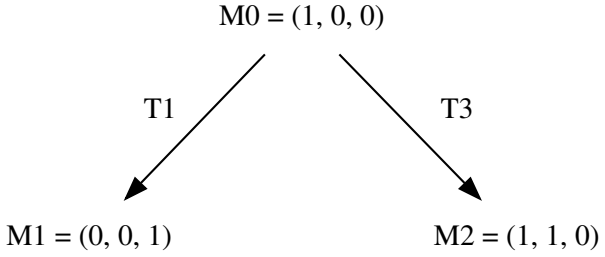

Figura 1.9: Primer paso para construir el árbol de alcanzabilidad de la red de Petri de la Fig. 1.8.

Ahora debemos considerar todas las marcas alcanzables desde las hojas del árbol.

A partir del marcado $(0, 0, 1)$ no podemos disparar ninguna transición. Esto se conoce como un *marcado muerto* (*dead marking*). En otras palabras se trata de un nodo "sin salida". Esta clase de estados finales es especialmente relevante para el análisis de bloqueos.

A partir de la marca de la derecha del árbol, denotada $(1, 1, 0)$, podemos disparar T1 o T3. Si disparamos T1, obtenemos $(0, 1, 1)$ y si dispara T3, la marca resultante es $(1, 2, 0)$. Esto produce el árbol de la Fig. 1.10.

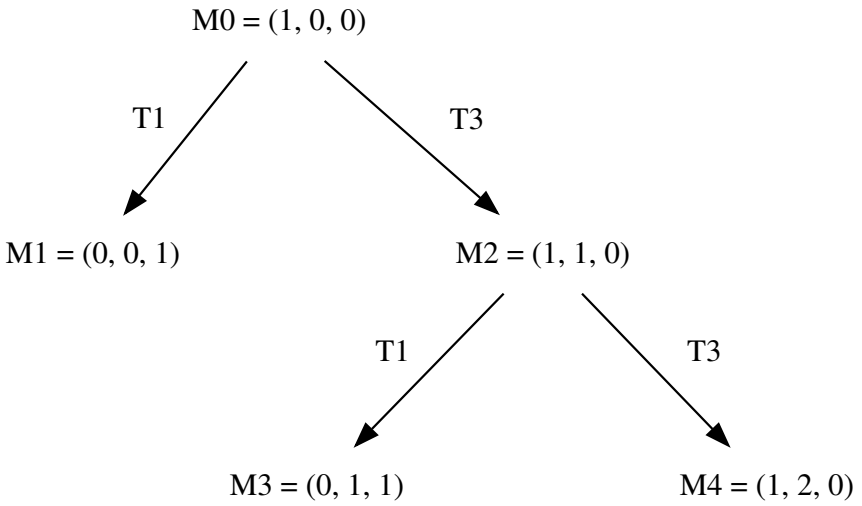

Figura 1.10: Segundo paso en la construcción del árbol de alcanzabilidad de la red de Petri de la Fig. 1.8.

Observe que partiendo de la marca $(0, 1, 1)$, sólo se habilita la transición T2 que conducirá a una marca $(0, 0, 1)$ ya vista anteriormente. Si en cambio tomamos $(1, 2, 0)$ tenemos de nuevo las mismas posibilidades que partiendo de $(1, 1, 0)$. Es fácil ver que el árbol seguirá creciendo por ese camino. Por tanto, el árbol es infinito y esto se debe a que la red de la Fig. 1.8 no está acotada. Veáse en la Fig. 1.11 el resultado final abreviado.

El método presentado anteriormente enumera los elementos del conjunto de alcanzabilidad. Se producirá cada marca del conjunto de alcanzabilidad y, por tanto, para cualquier red de Petri con un conjunto de alcanzabilidad infinito, es decir, un número infinito de estados posibles, el árbol correspondiente también sería infinito. Sin embargo, lo contrario no es cierto. Una red de Petri con un conjunto de alcanzabilidad finito puede tener un árbol infinito (véase la Fig. 1.12). Esta red es incluso *segura*. En conclusión, tratar con una red acotada o segura no es garantía de que el número total de estados alcanzables sea finito.

Para que el árbol de alcanzabilidad sea una herramienta de análisis útil, es necesario idear un método que lo limite a un tamaño finito. Esto implica en general una cierta pérdida de información, ya que el método tendrá que mapear (o mejor dicho reducir) un número infinito de marcados alcanzables a un solo elemento. La reducción a una representación finita puede lograrse por los siguientes medios.

Observe por un lado que podemos encontrarnos con nodos duplicados en nuestro árbol y que siempre los tratamos ingenuamente como nuevos. Esto se ilustra más claramente en la Fig. 1.12. Por tanto, es posible detener la exploración de los sucesores de un nodo duplicado.

Nótese, por otro lado, que algunas marcados son estrictamente diferentes de las marcas vistas anteriormente pero permiten el mismo conjunto de transiciones. Decimos en este caso que la

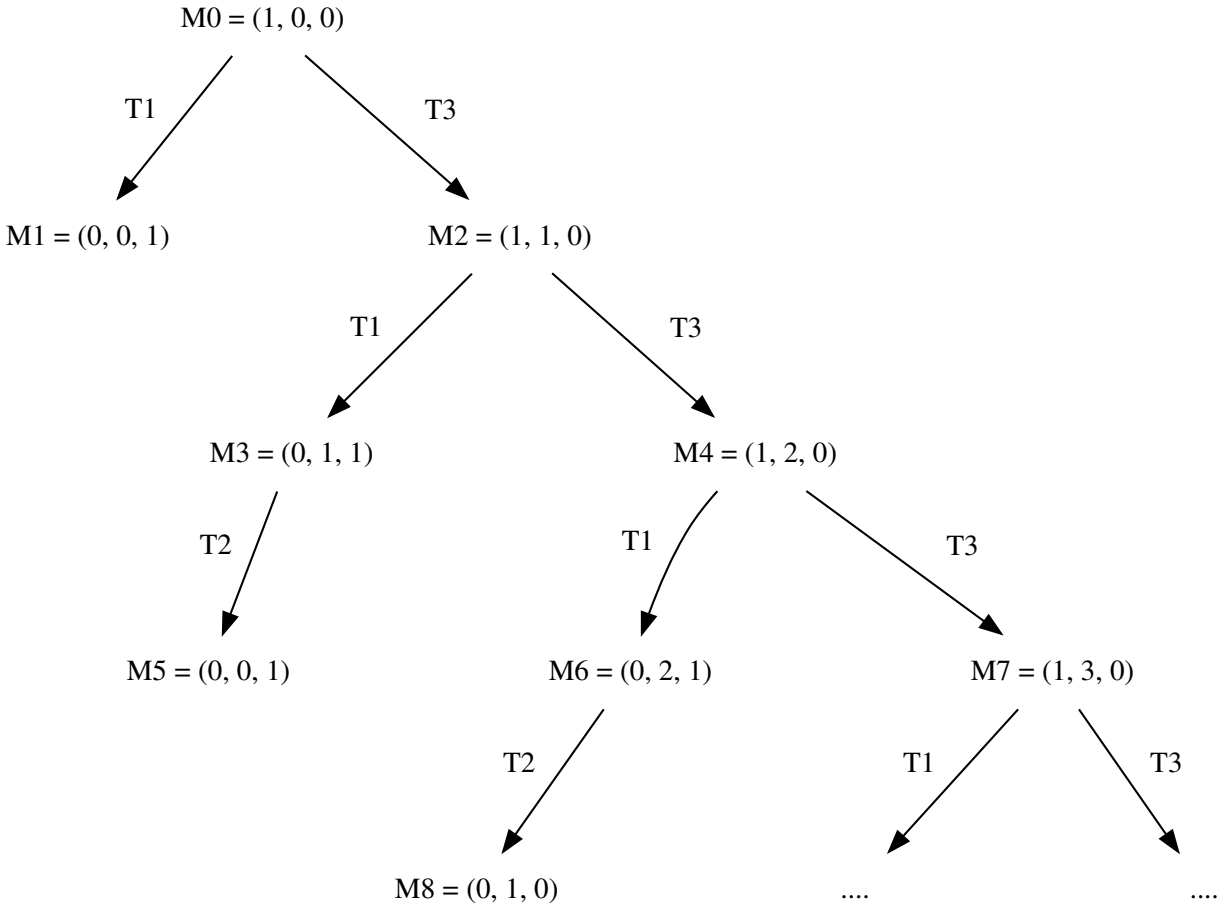

Figura 1.11: El árbol de alcanzabilidad infinita para la red de Petri de la Fig. 1.8.

marca con fichas adicionales *cubre* (*covers*) la que tiene el número mínimo de fichas necesarias para permitir el conjunto de transiciones en cuestión. Disparar algunas transiciones puede permitirnos acumular un número arbitrario de fichas en un lugar. Por ejemplo, disparar T3 en la red de Petri que se ve en la Fig. 1.8 muestra exactamente este comportamiento. En conclusión, bastaría con marcar el lugar de acumulación con una etiqueta especial ω, que significa infinito, ya que podríamos obtener tantas marcas como quisiéramos en ese lugar.

Por ejemplo, el resultado de convertir el árbol de la Fig. 1.11 en un árbol finito se muestra en la Fig. 1.13.

Para más detalles sobre

1. la técnica de representación de árboles de alcanzabilidad infinita mediante ω,
2. una definición del algoritmo y los pasos precisos para construir el árbol de alcanzabilidad,
3. la demostración matemática de que el árbol de alcanzabilidad generado por el algoritmo

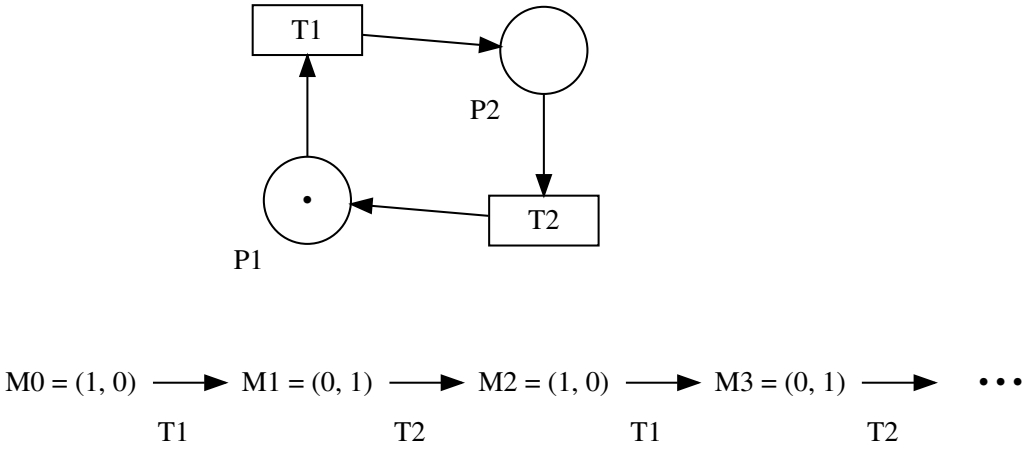

Figura 1.12: Una red de Petri simple con un árbol de alcanzabilidad infinito.

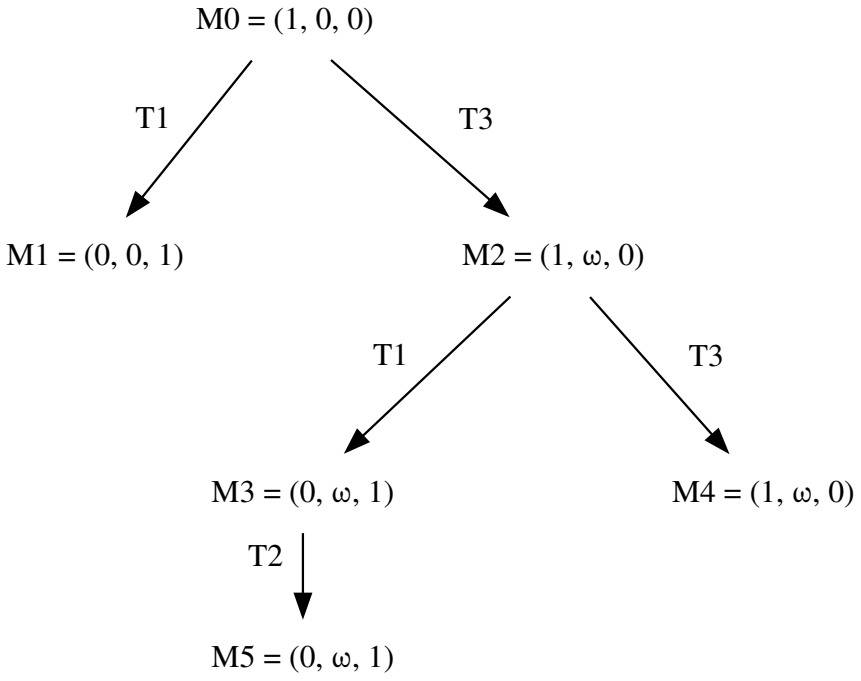

Figura 1.13: El árbol de alcanzabilidad finito para la red de Petri de la Fig. 1.8.

es finito,

4. y la distinción entre el árbol de alcanzabilidad y el *grafo de alcanzabilidad* (*reachability graph*)

se remite al lector a [Murata, 1989] y [Peterson, 1981]. Estos conceptos están fuera del alcance

de este trabajo y no son necesarios en los capítulos siguientes.

1.2. El lenguaje de programación Rust

Uno de los lenguajes de programación modernos más prometedores para la programación concurrente y segura para la memoria es Rust[1]. Rust es un lenguaje de programación multiparadigma y de propósito general cuyo objetivo es proporcionar a los desarrolladores una forma segura, concurrente y eficiente de escribir código de bajo nivel. Comenzó como un proyecto en Mozilla Research en 2009. La primera versión estable, Rust 1.0, se anunció el 15 de mayo de 2015. Para una breve historia de Rust hasta 2023, véase [Thompson, 2023].

El modelo de memoria de Rust basado en el concepto de propiedad (*ownership*) y su expresivo sistema de tipos evitan una gran variedad de clases de errores relacionados con la gestión de memoria y la programación concurrente en tiempo de compilación:

- Double free [Klabnik and Nichols, 2023, Chap. 4.1]
- Use after free [Klabnik and Nichols, 2023, Chap. 4.1]
- Dangling pointers [Klabnik and Nichols, 2023, Chap. 4.2]
- Data races [Klabnik and Nichols, 2023, Chap. 4.2] (con algunas advertencias importantes expuestas en [Rust Project, 2023c, Chap. 8.1])
- Pasar variables no seguras entre hilos [Klabnik and Nichols, 2023, Chap. 16.4]

El compilador oficial *rustc*[2] se encarga de controlar cómo se utiliza la memoria y de asignar y desasignar objetos. Si se encuentra una violación de sus reglas estrictas, el programa simplemente no será compilado.

En esta sección, justificaremos la elección de Rust para estudiar la detección de bloqueos y señales perdidas. Mostraremos cómo estos problemas pueden estudiarse por separado, sabiendo que otros errores ya se detectan en tiempo de compilación. En otras palabras, argumentaremos que la estabilidad y la seguridad del lenguaje proporcionan una base firme sobre la que construir una herramienta que detecte errores adicionales durante la compilación.

1.2.1. Características principales

Algunas de las principales características de Rust son:

- Sistema de tipos: Rust cuenta con un potente sistema de tipos que proporciona comprobaciones de seguridad en tiempo de compilación y evita muchos errores comunes de

[1] https://www.rust-lang.org/
[2] https://github.com/rust-lang/rust

programación. Incluye características como la inferencia de tipos, los tipos genéricos, los enums y el *pattern matching*. Cada variable tiene un tipo pero éste suele ser inferido por el compilador.

- Performance: La performance de Rust es comparable a la de C y C++ y a menudo es más rápido que muchos otros lenguajes de programación populares como Java, Go, Python o Javascript. La performance de Rust se debe a una combinación de características como las abstracciones de cero coste, un *runtime* mínimo y una gestión eficiente de la memoria.
- Concurrencia: Rust tiene un buen soporte por defecto para la concurrencia. Soporta varios paradigmas de concurrencia como el estado compartido, el pasaje de mensajes y la programación asíncrona. No obliga al desarrollador a implementar la concurrencia de una manera específica.
- *Ownership* y *borrowing*: Rust utiliza un modelo de *ownership* único para gestionar la memoria, lo que permite una asignación y desasignación de memoria eficientes sin riesgo de perder memoria o condiciones de carrera en el acceso a los datos. Además, no depende de un recolector de basura (*garbage collector*), ahorrando recursos. El verificador de préstamos (*borrow checker*) garantiza que sólo haya un propietario (*owner*) de un recurso en un momento dado.
- Impulsado por la comunidad: Rust cuenta con una vibrante y creciente comunidad de desarrolladores que contribuyen al desarrollo y al ecosistema del lenguaje. Cualquiera puede contribuir al lenguaje y sugerir mejoras. La documentación también es de código abierto y las decisiones importantes se documentan en forma de Requests for Comments (RFCs)[3].

El ciclo de publicación del compilador oficial de Rust, *rustc*, es notablemente rápido. Cada 6 semanas se publica una nueva versión estable del compilador [Klabnik and Nichols, 2023, Appendix G]. Esto es posible gracias a un complejo sistema de pruebas automatizado que compila incluso todos los paquetes disponibles en `crates.io`[4] utilizando un programa llamado *crater*[5] para verificar que la compilación y ejecución de las pruebas con la nueva versión del compilador no rompe los paquetes existentes [Albini, 2019].

El verificador de préstamos (*borrow checker*)

El verificador de préstamos (*borrow checker*) de Rust es un componente esencial de su modelo de *ownership*, diseñado para garantizar la seguridad de la memoria y evitar las carreras de datos (*data races*) en el código concurrente. El borrow checker analiza el código Rust en tiempo de compilación y aplica un conjunto de reglas para garantizar que se accede a la memoria de un programa de forma segura y eficiente.

[3] https://rust-lang.github.io/rfcs/
[4] https://crates.io/
[5] https://github.com/rust-lang/crater

La idea central detrás del borrow checker es que cada porción de memoria en un programa Rust tiene un propietario. El propietario puede cambiar durante la ejecución, pero sólo puede haber un propietario en un momento dado. Los valores de memoria también pueden tomarse *prestados* (*borrowed*), es decir, utilizarse sin cambiar el propietario, de forma similar al acceso al valor a través de un puntero o una referencia en otros lenguajes de programación. Cuando se toma prestado un valor, el prestatario recibe una referencia al valor, pero el propietario original conserva la propiedad. El verificador de préstamos aplica reglas para garantizar que un valor prestado no se modifica mientras está prestado y que el prestatario libera la referencia antes de que el propietario salga de *scope*.

Para mayor claridad, presentaremos a continuación algunas de las reglas centrales aplicadas por el borrow checker:

- No pueden existir simultáneamente dos referencias mutables a la misma posición de memoria. Esto evita las carreras de datos en las que dos hilos intentan modificar la misma posición de memoria al mismo tiempo.
- Las referencias mutables no pueden existir al mismo tiempo que las referencias inmutables a la misma ubicación de memoria. Esto garantiza que las referencias mutables e inmutables no puedan utilizarse simultáneamente, evitando lecturas y escrituras incoherentes.
- Las referencias no pueden sobrevivir al valor al que hacen referencia. Esto garantiza que las referencias no apunten a ubicaciones de memoria no válidas, evitando desferencias de punteros nulos y otros errores de memoria.
- Las referencias no pueden utilizarse después de que su propietario haya sido desplazado o destruido. Esto asegura que las referencias no apunten a memoria que ha sido desasignada, evitando errores de uso después de liberar.

Puede requerir cierto esfuerzo escribir código Rust que satisfaga estas reglas. El borrow checker suele señalarse como un aspecto del lenguaje que resulta confuso para los nuevos usuarios. Sin embargo, esta disciplina adquirida paga sus frutos en términos de mayor seguridad de memoria y performance durante la ejecución. Al garantizar que los programas Rust siguen estas reglas, el verificador de préstamos elimina muchos errores comunes de programación que pueden dar lugar a fugas de memoria, *data races* y otros bugs, al tiempo que enseña buenas prácticas y patrones de programación.

Manejo de errores aplicado por el compilador

La gestión de errores es un aspecto esencial de la programación y suele abordarse en el diseño de los lenguajes de programación. La miríada de enfoques puede resumirse en dos grupos distintos.

Un grupo formado por lenguajes como C++, Java o Python emplea excepciones, utilizando bloques `try` y `catch` para manejar condiciones excepcionales. Cuando se lanzan excepciones y no se capturan, el programa termina abruptamente.

El otro grupo lo forman lenguajes como C o Go, entre otros, en los que la convención es comunicar un error a través del valor de retorno de las funciones o mediante un parámetro de función específicamente dedicado a este fin. La desventaja es que el compilador no impone al programador la comprobación de errores, lo que puede hacer que no se tengan en cuenta los casos de error al añadir nuevas funciones.

Rust adopta un enfoque diferente promoviendo la noción de que las funciones idealmente no deberían fallar y que la firma de la función debería reflejar si la función puede devolver un error. En lugar de excepciones o códigos de error con números enteros, las funciones Rust que pueden terminar con errores devuelven un tipo `std::result::Result`[6] que puede contener el resultado del cálculo o un tipo de error personalizado acompañado de una descripción del error. *rustc* impone que el programador escriba código para ambos casos y el lenguaje proporciona mecanismos para facilitar la gestión de errores [Klabnik and Nichols, 2023, Chap. 9.2].

En Rust, el foco está puesto en la gestión coherente del caso de error. Los errores pueden propagarse a las llamadas a funciones de nivel superior hasta que pueda restablecerse un estado coherente del programa. Sin embargo, puede haber situaciones en las que la recuperación de un estado de error no sea factible. En tales casos, se puede ordenar al programa que entre en pánico, lo que resulta en un cierre abrupto y sin gracia (*ungraceful*), similar a una excepción no capturada en otros lenguajes de programación. Durante un pánico, la ejecución del programa se aborta y la pila se despliega (*stack unwinding*) [Klabnik and Nichols, 2023, Chap. 9.1]. Se genera un mensaje de error que contiene detalles del pánico, por ejemplo, el propio mensaje de error y su ubicación. Aunque los *panic* pueden ser capturados por hilos padre (*parent threads*) y en casos específicos cuando el programador así lo desea[7], normalmente conducen a la terminación del programa actual. Este mecanismo de pánico estructurado hace que el compilador sea consciente de los posibles errores irrecuperables, lo que permite la generación del código adecuado para manejar estos casos.

Rust también proporciona un tipo `std::option::Option`[8] que representa tanto la presencia de un valor como su ausencia. De nuevo, el compilador impone disciplina al programador para manejar siempre el caso `None`. De este modo, Rust elimina casi por completo la necesidad de un puntero NULL como se encuentra en otros lenguajes como C, C++, Java, Python o Go.

1.2.2. Adopción

En esta subsección, describiremos brevemente la tendencia en la adopción del lenguaje de programación Rust. Esto resalta la relevancia de este trabajo como contribución a una comunidad creciente de programadores que hacen hincapié en la importancia de una programación de sistemas segura y eficaz para los próximos años en la industria del software.

[6] https://doc.rust-lang.org/std/result/
[7] https://doc.rust-lang.org/std/panic/fn.catch_unwind.html
[8] https://doc.rust-lang.org/std/option/

En los últimos años, varios proyectos importantes de la comunidad de código abierto y de empresas privadas han decidido incorporar Rust para reducir el número de bugs relacionados con la gestión de la memoria sin sacrificar performance. Entre ellos, podemos citar algunos ejemplos representativos:

- El Android Open Source Project fomenta el uso de Rust para los componentes del SO por debajo del Android Runtime (ART) [Stoep and Hines, 2021].
- El kernel Linux introduce en la versión 6.1 (publicada en diciembre de 2022) soporte oficial de herramientas para la programación de componentes en Rust [Corbet, 2022, Simone, 2022].
- En Mozilla, el proyecto Oxidation se creó en 2015 para aumentar el uso de Rust en Firefox y proyectos relacionados. En marzo de 2023, las líneas de código en Rust representan más del 10 % del total en Firefox Nightly [Mozilla Wiki, 2015].
- En Meta, el uso de Rust como lenguaje de desarrollo del lado del servidor está aprobado y es alentado desde julio de 2022 [Garcia, 2022].
- En Cloudflare, se construyó desde cero un nuevo proxy HTTP en Rust para superar las limitaciones architectónicas de NGINX, reduciendo el uso de CPU en un 70 % y el de memoria en un 67 % [Wu and Hauck, 2022].
- En Discord, la reimplementación en Rust de un servicio crucial escrito en Go proporcionó grandes beneficios en el rendimiento y resolvió una pérdida de rendimiento debida al *garbage collection* en Go [Howarth, 2020].
- En npm Inc, la empresa detrás del npm registry, Rust permitió escalar los servicios limitados por la cantidad de CPU disponible a más de 1.300 millones de descargas al día [The Rust Project Developers, 2019].

En otros casos, Rust ha demostrado ser una gran elección en proyectos C/C++ existentes para reescribir módulos que procesan entradas de usuario no fiables, por ejemplo, analizadores sintácticos, y reducir el número de vulnerabilidades de seguridad debidas a problemas de memoria [Chifflier and Couprie, 2017].

Además, el interés de la comunidad de desarrolladores por Rust es innegable, ya que ha sido calificado durante 7 años consecutivos como el lenguaje de programación más "querido"(*loved*) por los programadores en la encuesta Stack Overflow Developer Survey [Stack Overflow, 2022].

1.2.3. Importancia del uso seguro de la memoria

En esta subsección, se presentan pruebas convincentes que apoyan el uso de un lenguaje de programación que soporte un uso seguro de la memoria. El objetivo es resaltar la importancia de avanzar en la investigación sobre la detección de errores en tiempo de compilación para evitar fallos que posteriormente son difíciles de corregir en los sistemas en producción.

Varias investigaciones empíricas han concluido que alrededor del 70 % de las vulnerabilidades encontradas en grandes proyectos C/C++ se deben a errores en el manejo de la memoria. Esta cifra elevada puede observarse en proyectos como:

- Android Open Source Project [Stepanov, 2020],
- los componentes Bluetooth y multimedia de Android [Stoep and Zhang, 2020],
- el Proyecto Chromium detrás del navegador web Chrome [The Chromium Projects, 2015],
- el componente CSS de Firefox [Hosfelt, 2019],
- iOS y macOS [Kehrer, 2019],
- productos de Microsoft [Miller, 2019, Fernandez, 2019],
- Ubuntu [Gaynor, 2020]

Numerosas herramientas se han fijado el objetivo de abordar estas vulnerabilidades causadas por una asignación inadecuada de memoria en bases de código ya establecidas. Sin embargo, su uso conlleva una pérdida notable de rendimiento y no todas las vulnerabilidades pueden evitarse [Szekeres et al., 2013]. Un ejemplo de herramienta representativa en este ámbito, más concretamente un detector dinámico de data races para programas multihilo en C, puede encontrarse en [Savage et al., 1997], cuyo algoritmo se mejoró posteriormente en [Jannesari et al., 2009] y se integró en la herramienta Helgrind, parte del conocido framework de instrumentación Valgrind[9].

En [Jaeger and Levillain, 2014], los autores ofrecen un estudio detallado de las características de los lenguajes de programación que comprometen la seguridad de los programas resultantes. Hablan de las características de seguridad intrínsecas de los lenguajes de programación y enumeran recomendaciones para la formación de desarrolladores o evaluadores de software seguro. La seguridad de tipos se menciona como uno de los elementos clave para eliminar clases completas de errores desde el principio. Otra consideración digna de mención es utilizar un lenguaje en el que las especificaciones sean lo más completas, explícitas y formalmente definidas posible. El concepto de Undefined Behavior (UB) debe incluirse con precaución y sólo con moderación. Algunos ejemplos de la especificación C/C++ ilustran la confusión que se deriva de no seguir estos principios. Los autores concluyen que la seguridad de la memoria conseguida mediante la recogida de basura (*garbage collection*) supone una amenaza para la seguridad y que en su lugar deberían considerarse otros mecanismos.

Debemos tener en cuenta que el propio Rust, como cualquier otra producto de software, no está exento de vulnerabilidades de seguridad. En el pasado se han descubierto bugs serios en la biblioteca estándar [Davidoff, 2018]. Además, la generación de código en Rust también incluye mitigaciones a exploits de diversa índole [Rust Project, 2023b, Chap. 11]. Sin embargo, esto dista mucho de los problemas ampliamente conocidos en C y C++.

[9] `https://valgrind.org/`

1.3. Correctitud de programas concurrentes

En el área de la computación concurrente, uno de los principales retos es demostrar la correctitud de un programa concurrente. A diferencia de un programa secuencial en el que para cada entrada se obtiene siempre la misma salida, en un programa concurrente la salida puede depender de cómo se hayan intercalado las instrucciones de los distintos procesos o hilos durante la ejecución.

La correctitud de un programa concurrente se define entonces en términos de las propiedades de la computación realizada y no sólo en términos del resultado obtenido. En la literatura [Ben-Ari, 2006, Coulouris et al., 2012, van Steen and Tanenbaum, 2017], se definen dos tipos de propiedades de correctitud:

- **Propiedades de seguridad (*Safety properties*):**: La propiedad debe ser *siempre* verdadera.
- **Propiedades de liveness**: La propiedad debe volverse *eventualmente* verdadera.

Dos propiedades de seguridad deseables en un programa concurrente son:

- **Exclusión mutua**: Dos procesos no deben acceder a los recursos compartidos al mismo tiempo.
- **Ausencia de bloqueo (*deadlock*)**: Un sistema en funcionamiento debe poder seguir realizando su tarea, es decir, progresando y produciendo trabajo útil.

Las primitivas de sincronización como los mutexes, los monitores (propuestos por [Hansen, 1972, Hansen, 1973]), los semáforos (propuestos por [Dijkstra, 2002]) y las variables de condición (propuestas por [Hoare, 1974]) suelen utilizarse para implementar el acceso coordinado de hilos o procesos a recursos compartidos. Sin embargo, el uso correcto de estas primitivas es difícil de conseguir en la práctica y puede introducir errores difíciles de detectar y corregir. Actualmente, la mayoría de los lenguajes de propósito general, ya sean compilados o interpretados, no permiten detectar estos errores en todos los casos.

Dada la creciente importancia de la programación concurrente debido a la proliferación de sistemas de hardware multihilo y multihilo, minimizar la aparición de errores asociados a la sincronización de hilos o procesos tiene una importancia innegable para la industria. El funcionamiento libre de deadlocks es un requisito fundamental para muchos proyectos, como los sistemas operativos [Arpaci-Dusseau and Arpaci-Dusseau, 2018], las aeronaves [Carreño and Muñoz, 2005, Monzon and Fernandez-Sanchez, 2009] y los vehículos autónomos [Perronnet et al., 2019].

En la próxima sección examinaremos más detenidamente las condiciones que provocan un deadlock y las estrategias utilizadas para hacerles frente.

1.4. Bloqueo mutuo (*deadlocks*)

Los bloqueos o *deadlocks* son un problema común en sistemas concurrentes, es decir, sistemas en los que varios hilos o procesos se ejecutan simultáneamente y potencialmente comparten recursos. Se han estudiado al menos desde [Dijkstra, 1964] quien acuñó el término "abrazo mortal" en holandés el cual cayó eventualmente en desuso.

Un deadlock se produce cuando dos o más hilos o procesos están bloqueados y no pueden seguir ejecutándose porque cada uno está esperando a que el otro libere un recurso que necesita. Esto da lugar a una situación en la que ninguno de los hilos o procesos puede progresar y el sistema queda efectivamente atascado. En [Holt, 1972] puede encontrarse una definición alternativa equivalente de los bloqueos en términos de los estados del programa.

Los deadlocks pueden ser un problema grave en los sistemas concurrentes, ya que pueden provocar que el sistema deje de responder o incluso que se interrumpa la ejecución abruptamente. Por lo tanto, sería ventajoso poder detectar y prevenir los dealocks. Pueden producirse en cualquier sistema concurrente en el que varios hilos o procesos compiten por recursos compartidos. Algunos ejemplos de recursos compartidos que pueden provocar deadlocks son la memoria del sistema, los dispositivos de entrada/salida, los *locks* y otros tipos de primitivas de sincronización.

Los bloqueos pueden ser difíciles de detectar y prevenir porque dependen de la sincronización precisa de los eventos en el sistema. Incluso en los casos en los que pueden detectarse los deadlocks, resolverlos puede ser difícil, ya que puede requerir liberar recursos que ya han sido adquiridos o deshacer transacciones completadas. Para evitar los bloqueos, es importante gestionar cuidadosamente los recursos compartidos en un sistema concurrente. Esto puede implicar el uso de técnicas como algoritmos de asignación de recursos, algoritmos de detección de bloqueos y otros tipos de primitivas de sincronización. Gestionando cuidadosamente los recursos compartidos es posible evitar que se produzcan bloqueos y garantizar el buen funcionamiento de los sistemas concurrentes.

Para entender el concepto con más detalle, considérese un ejemplo sencillo en el que dos procesos, A y B, compiten por dos recursos, X e Y. Inicialmente, el proceso A ha adquirido el recurso X y está esperando adquirir el recurso Y, mientras que el proceso B ha adquirido el recurso Y y está esperando adquirir el recurso X. En esta situación, ninguno de los dos procesos puede seguir ejecutándose porque está esperando a que el otro proceso libere un recurso que necesita. Esto da lugar a un deadlock, ya que ninguno de los dos procesos puede progresar. La Fig. 1.14 ilustra esta situación. El ciclo que aparece en ella indica un deadlock como se explicará en la siguiente sección.

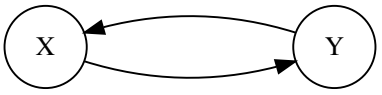

Figura 1.14: Ejemplo de un grafo de estados con un ciclo que indica un bloqueo mutuo.

1.4.1. Condiciones necesarias

Según el paper clásico sobre el tema [Coffman et al., 1971], deben darse las siguientes condiciones para que se produzca un deadlock. A veces se las denominan "condiciones de Coffman".

1. **Exclusión mutua (*Mutual Exclusion*)**: Al menos un recurso del sistema debe mantenerse en un modo no compartible, lo que significa que sólo un hilo o proceso puede utilizarlo a la vez, por ejemplo, una variable detrás de un mutex.
2. **Retener y esperar (*Hold and Wait*)**: Al menos un hilo o proceso del sistema debe estar reteniendo un recurso y esperando para adquirir recursos adicionales que en ese momento están siendo retenidos por otros hilos o procesos.
3. **Sin apropiación (*No Preemption*)**: Los recursos no pueden apropiarse, lo que significa que un hilo o proceso que posea un recurso no puede ser obligado a liberarlo hasta que haya completado su tarea.
4. **Espera circular (*Circular Wait*)**: Debe haber una cadena circular de dos o más hilos o procesos, en la que cada hilo o proceso esté esperando un recurso en poder del siguiente de la cadena. Esto suele visualizarse en un gráfico que representa el orden en que se adquieren los recursos.

Usualmente, las tres primeras condiciones son características del sistema estudiado, es decir, los protocolos utilizados para adquirir y liberar recursos, mientras que la cuarta puede materializarse o no en función del intercalado de instrucciones durante la ejecución.

Cabe señalar que las condiciones de Coffman son en general necesarias pero no suficientes para que se manifieste un bloqueo. En efecto, las condiciones son suficientes en el caso de sistemas de recursos de una sola instancia (una unidad de cada recurso). Pero sólo indican la posibilidad de un deadlock en los sistemas en los que hay múltiples instancias indistinguibles del mismo recurso.

En el caso general, si no se cumple alguna de las condiciones, no puede producirse un bloqueo, pero la presencia de las cuatro condiciones no garantiza necesariamente un bloqueo. No obstante, las condiciones de Coffman son un marco útil para comprender y analizar las causas de los bloqueos en los sistemas concurrentes y pueden ayudar a orientar el desarrollo de estrategias para prevenir y resolver los bloqueos.

1.4.2. Estrategias

Existen varias estrategias para gestionar los bloqueos, cada una de las cuales tiene sus puntos fuertes y débiles. En la práctica, la estrategia más eficaz dependerá de los requisitos y limitaciones específicos del sistema que se esté desarrollando. Los diseñadores y desarrolladores deben considerar cuidadosamente las compensaciones entre las distintas estrategias y

elegir el enfoque que mejor se adapte a sus necesidades. Se remite a los lectores interesados a [Coffman et al., 1971, Singhal, 1989].

Prevención

Una forma de hacer frente a los bloqueos es evitar que se produzcan en primer lugar. La idea es que los bloqueos se excluyan a priori. Con este objetivo en mente, debemos asegurarnos de que en cada momento no se cumple al menos una de las condiciones necesarias desarrolladas en la Sec. 1.4.1. Esto restringe los posibles protocolos en los que se pueden realizar solicitudes de recursos. A continuación examinaremos cada condición por separado y desarrollaremos los enfoques más comunes.

Si la primera condición debe ser falsa, entonces el programa debe permitir el acceso compartido a todos los recursos. Los algoritmos de sincronización sin bloqueo pueden utilizarse para este fin, ya que no implementan la exclusión mutua. Esto es difícil de conseguir en la práctica para todos los tipos de recursos, ya que, por ejemplo, un archivo no puede ser compartido por más de un hilo o proceso durante una actualización del contenido del archivo.

En cuanto a la segunda condición, un enfoque viable sería imponer que cada hilo o proceso adquiera todos los recursos necesarios a la vez y que el hilo o proceso no pueda continuar hasta que se le haya concedido acceso a todos ellos. Esta política de "todo o nada" provoca una penalización significativa del rendimiento, dado que los recursos pueden asignarse a un hilo o proceso específico pero pueden permanecer sin utilizar durante largos periodos. En términos más sencillos, disminuye la concurrencia.

Si se deniega la condición de no apropiación, los recursos pueden recuperarse en determinadas circunstancias, por ejemplo, utilizando algoritmos de asignación de recursos que garanticen que los recursos nunca se retienen indefinidamente. Tras un tiempo de espera o cuando se cumple una condición, el hilo o proceso libera el recurso o un proceso supervisor lo recupera a la fuerza. Normalmente, esto funciona bien cuando el estado del recurso puede guardarse fácilmente y restaurarse más tarde. Un ejemplo de ello es la asignación de núcleos de CPU en un sistema operativo (OS) moderno. El *scheduler* asigna un núcleo de procesador a una tarea y puede cambiar a una tarea diferente o puede mover la tarea a un nuevo núcleo de procesador en cualquier momento simplemente guardando el contenido de los registros [Arpaci-Dusseau and Arpaci-Dusseau, 2018, Chap. 6]. Sin embargo, si no es posible preservar el estado de los recursos, el adelantamiento puede implicar una pérdida del progreso realizado hasta el momento, lo que no es aceptable en muchos escenarios.

Por último, si el grafo de estados de los recursos nunca forma un ciclo, entonces la cuarta condición necesaria es falsa y se evitan los bloqueos. Para lograrlo, se podría introducir una ordenación lineal de los tipos de recursos. En otras palabras, si a un proceso o hilo se le han asignado recursos del tipo r_i, podrá requerir posteriormente sólo aquellos recursos de tipos que sigan a r_i en el ordenamiento. Esto implica utilizar primitivas de sincronización especiales que permitan compartir recursos de forma controlada y aplicar reglas estrictas para la adquisición

y liberación de recursos. En estas condiciones, el grafo de estado será estrictamente un bosque (un grafo acíclico), por lo que no es posible que se produzcan bloqueos.

En aplicaciones prácticas, una combinación de las estrategias anteriores puede resultar útil cuando ninguna de ellas sea totalmente aplicable.

Evasión (*avoidance*)

La evitación es otra estrategia para hacer frente a los bloqueos, que consiste en detectar y evitar dinámicamente los bloqueos potenciales *antes* de que se produzcan. Para ello, el sistema requiere un conocimiento global por adelantado sobre qué recursos solicitará un hilo o proceso durante su vida. Tenga en cuenta que, en términos lingüísticos, "evitar un bloqueo" y "prevenir un bloqueo" pueden parecer similares, pero en el contexto de la gestión de bloqueos, son conceptos distintos.

Uno de los algoritmos clásicos para evitar el bloqueo es el algoritmo de Banker [Dijkstra, 1964]. Otro algoritmo relevante es el propuesto por [Habermann, 1969].

Lamentablemente, estas técnicas sólo son efectivas en escenarios muy específicos, como en un sistema embebido en el que se conoce a priori el conjunto completo de tareas a ejecutar y el *locks* necesarios. En consecuencia, la evitación de bloqueos no es una solución de uso común aplicable a una amplia gama de situaciones.

Detección y recuperación

Otra estrategia para gestionar los bloqueos es detectarlos *después* de que se produzcan y recuperarse de ellos. Para un estudio de los algoritmos de detección de bloqueos en sistemas distribuidos, véase [Singhal, 1989]. Presentaremos brevemente la idea general que subyace a uno de ellos con fines ilustrativos.

El gráfico de asignación de recursos (Resource Allocation Graph (RAG)) es un método comúnmente utilizado para detectar bloqueos en sistemas concurrentes. Representa la relación entre hilos/procesos y recursos en el sistema como un grafo dirigido. Cada proceso y recurso está representado por un nodo en el grafo y se traza una arista dirigida desde un proceso a un recurso si el proceso está actualmente ocupando ese recurso. Esto es análogo al grafo de estado mostrado en la Fig. 1.14 pero con los hilos/procesos representados en el diagrama. El grafo de estado también puede aplicarse a la detección de bloqueos [Coffman et al., 1971].

Para detectar bloqueos mediante el RAG tenemos que buscar ciclos en el gráfico. Si hay un ciclo en el gráfico, indica que un conjunto de procesos está esperando recursos que en ese momento están en manos de otros procesos del ciclo. Por lo tanto, ningún proceso del ciclo puede avanzar.

La parte de recuperación del proceso consiste en terminar uno de los hilos o procesos del ciclo. Esto hace que se liberen los recursos y que los demás hilos o procesos puedan continuar.

Los sistemas de gestión de bases de datos (Database management systems (DBMS)) incorporan subsistemas para detectar y resolver los bloqueos. Un detector de bloqueos se ejecuta a intervalos, generando un gráfico de asignación regular, también llamado gráfico de transacción-espera (transaction-wait-for (TWF)), y examinándolo en busca de cualquier ciclo. Si se identifica un ciclo (deadlock), el sistema debe reiniciarse. Una excelente visión general de la detección de bloqueos en sistemas de bases de datos distribuidos es [Knapp, 1987]. El tema del control de la concurrencia y la recuperación de los bloqueos en los DBMS se trata ampliamente en [Bernstein et al., 1987].

Aceptar o ignorar por completo los deadlocks

En algunos casos, puede ser admisible aceptar simplemente el riesgo de que se produzcan deadlocks y gestionarlos a medida que vayan apareciendo. Este enfoque puede ser adecuado en sistemas en los que el coste de prevenir o detectar los deadlocks sea demasiado elevado, o en los que la frecuencia de los deadlocks sea lo suficientemente baja como para que el impacto en el rendimiento del sistema sea mínimo, o en los que la pérdida de datos que se produzca cada vez sea tolerable.

UNIX es un ejemplo de sistema operativo que sigue este principio [Shibu, 2016, p. 477]. Otros sistemas operativos populares también muestran este comportamiento. Por otro lado, un sistema de misión crítica no puede permitirse fingir que su funcionamiento estará libre de bloqueos por ningún motivo.

1.5. Condition variables

Las variables de condición (*condition variables*) son una primitiva de sincronización en la programación concurrente que permite a los hilos esperar eficientemente a que se cumpla una condición específica antes de continuar. Fueron introducidas por primera vez por [Hoare, 1974] como parte de un bloque de construcción para el concepto de monitor desarrollado originalmente por [Hansen, 1973].

Siguiendo la definición clásica, se pueden llamar dos operaciones principales sobre una variable de condición:

- `esperar (wait)`: Bloquea el hilo o proceso actual. En algunas implementaciones, el mutex asociado se libera como parte de la operación.
- `señal (signal)`: Despierta un hilo o proceso que espera en la condition variable. En algunas implementaciones, el mutex asociado es adquirido inmediatamente por el hilo o proceso sobre el que se aplica la operación.

Las condition variables suelen estar asociadas a un predicado booleano (una condición) y a un mutex. El predicado booleano es la condición que esperan los hilos o procesos. Cuando

se establece en un valor determinado (verdadero o falso), el hilo o proceso puede continuar ejecutándose. El mutex garantiza que sólo un hilo o proceso pueda acceder a la condition variable a la vez.

Las variables de condición no contienen un valor real accesible para el programador en su interior. En su lugar, se implementan utilizando una estructura de datos de cola donde los hilos o procesos se añaden a la cola cuando entran en el estado de espera. Cuando otro hilo o proceso señala el estado, se selecciona un elemento de la cola para reanudar la ejecución. La política de *scheduling* específica puede variar en función de la implementación.

A lo largo de los años, se han desarrollado diversas implementaciones y optimizaciones para las condition variables con el fin de mejorar el rendimiento y reducir la sobrecarga. Por ejemplo, algunas implementaciones permiten despertar varios hilos a la vez (una operación denominada *broadcast*), mientras que otras utilizan una cola de prioridad para lograr que los hilos de mayor prioridad se despierten primero.

Las condition variables forman parte de la biblioteca estándar POSIX para hilos (`pthreads`) [Nichols et al., 1996] y en la actualidad se utilizan ampliamente en lenguajes y sistemas de programación concurrentes. Se encuentran entre otros en:

- UNIX[10],
- Rust[11]
- Python[12]
- Go[13]
- Java[14]

A pesar de su uso generalizado, las variables de condición pueden ser difíciles de utilizar correctamente, y un uso incorrecto puede dar lugar a errores sutiles y difíciles de depurar, como señales perdidas o despertares espurios. A continuación veremos estos errores en detalle.

1.5.1. Señales perdidas

Una señal perdida ocurre cuando un hilo o proceso que espera en una condition variable no recibe una señal aunque haya sido emitida. Esto puede ocurrir debido a una condición de carrera en la que la señal se emite antes de que el hilo entre en estado de espera, provocando que se pierda la señal.

[10] https://man7.org/linux/man-pages/man3/pthread_cond_init.3p.html
[11] https://doc.rust-lang.org/std/sync/struct.Condvar.html
[12] https://docs.python.org/3/library/threading.html
[13] https://pkg.go.dev/sync
[14] https://docs.oracle.com/en/java/javase/20/docs/api/java.base/java/util/concurrent/locks/Condition.html

Para ilustrar el concepto de señal perdida, veremos un ejemplo. Supongamos que tenemos dos hilos, T1 y T2, y una variable entera compartida llamada `flag`. T1 establece `flag` en `true` y envía una señal a una variable de condición `cv` para despertar a T2 que está esperando en cv para saber cuándo se ha marcado `flag`. T2 espera en `cv` hasta que recibe una señal de T1. El listado 1.1 muestra el pseudocódigo correspondiente.

```
1  // T1
2  lock.acquire()
3  flag = true
4  cv.signal()   // Signal T2 to wake up
5  lock.release()
6
7  // T2
8  lock.acquire()
9  while (flag == false)   // Wait until flag has changed
10     cv.wait(lock)
11 lock.release()
```

Listing 1.1: Pseudocódigo para un ejemplo de señal perdida.

Supongamos ahora que T1 activa `flag` y emite una señal a `cv` pero T2 aún no ha entrado en estado de espera en `cv` debido a algún retraso en *scheduling*. En este caso, la señal emitida por T1 podría ser pasada por alto por T2, como se muestra en la siguiente secuencia de acontecimientos:

1. T1 adquiere el bloqueo y marca `flag` como `true`.
2. T1 indica a `cv` que despierte a T2.
3. T1 libera el lock.
4. T2 adquiere el bloqueo y comprueba si `flag` ha cambiado. Como `flag` sigue siendo `flag`, T2 entra en estado de espera en `cv`.
5. Debido a retrasos en el *scheduling* o a otros factores, T2 no recibe la señal emitida por T1 y permanece atascado en el estado de espera para siempre.

Este escenario ilustra el concepto de señal perdida, en el que un hilo que espera en una variable de condición no recibe una señal aunque haya sido emitida. Para evitar que se pierdan señales, es esencial asegurarse de que los hilos que esperan en variables de condición estén correctamente sincronizados con los hilos que emiten señales y de que no existan condiciones de carrera o problemas de sincronización que puedan hacer que se pierdan señales.

1.5.2. Despertares espurios (*spurious wakeups*)

Un despertar espurio (*spurious wakeup*) ocurre cuando un hilo que espera en una variable de condición se despierta sin recibir una señal o notificación de otro hilo. Las razones son múltiples: interrupciones del hardware o del sistema operativo, detalles internos de implementación de la condition variable u otros factores impredecibles.

Reutilizando la situación descrita en la sección anterior y el pseudocódigo mostrado en el Listado 1.1, supongamos ahora que T1 pone el `flag` em `true` y emite una señal a `cv`, pero T2 se despierta sin recibir la señal emitida por T1.

Este es precisamente el despertar espurio. La siguiente secuencia de acontecimientos conduce a este desafortunado resultado:

1. T1 adquiere el bloqueo y marca `flag` como `true`.
2. T1 indica a `cv` que despierte a T2.
3. T1 libera el lock.
4. T2 adquiere el lock y comprueba si `flag` es verdadero. Como `flag` sigue siendo `false`, T2 entra en estado de espera en cv.
5. Debido a algún detalle de implementación interna de la condition variable o a otros factores impredecibles, T2 se despierta sin recibir la señal emitida por T1 y continúa ejecutando la siguiente sentencia de su código.

Este ejemplo demuestra la idea de un despertar espurio en el que un hilo que espera en una condition variable se despierta sin recibir una señal o notificación de otro hilo. Para evitar los despertares espurios es inevitable utilizar un bucle para volver a comprobar la condición después de despertarse de un estado de espera, como se muestra en el pseudocódigo para T2 (línea 9). Esto garantiza que el hilo no prosiga hasta que la condición que está esperando se haya producido realmente. Si no existiera el bucle while, un despertar espurio haría que T2 continuara ejecutándose después de la llamada a `wait`, independientemente de si T1 emitió una señal o no.

1.6. Arquitectura del compilador

Los compiladores son programas que transforman el código fuente escrito en un lenguaje en otro lenguaje, normalmente código máquina. Un compilador toma un programa en un lenguaje, el *lenguaje fuente*, y lo traduce a un programa equivalente en otro lenguaje, el *lenguaje de destino*.

Para lograrlo, los compiladores suelen tener una serie de fases o pases que se ejecutan en secuencia. El objetivo de estos pases es traducir el código de alto nivel en código de bajo nivel que la máquina pueda ejecutar. Tras cada pasada, el código se acerca cada vez más a la

representación final. Estas fases están hoy en día bien definidas y diferentes compiladores las implementan en alguna forma u otra [Aho et al., 2014, Chap. 1.2].

La primera pasada de un compilador típico es la fase de **análisis léxico (*lexical analysis*)**. En esta fase, el código fuente se descompone en un flujo de tokens, cada uno de los cuales representa una única pieza del código. El analizador léxico (*lexer*) identifica palabras clave, identificadores, literales y otros tokens que forman los bloques de construcción del código fuente.

La siguiente pasada es la fase de **análisis sintáctico (*syntax analysis*)**, también conocida como fase del *parser*. En esta fase, los tokens producidos por el *lexer* se analizan según las reglas de la gramática del lenguaje de programación. El analizador sintáctico construye un *parse tree* o un árbol sintáctico abstracto (abstract syntax tree (AST)) que representa la estructura del código.

La tercera pasada es la fase de **análisis semántico (*semantic analysis*)**, en la que el compilador comprueba la correctitud semántica del código, como la comprobación de errores de tipo, variables no definidas y operaciones no válidas. El analizador semántico (*semantic analyzer*) construye una tabla de símbolos que contiene información sobre las variables, funciones y otras entidades definidas en el código.

La cuarta pasada es la fase de **generación de código (*code generation*)**. El compilador toma el AST y la tabla de símbolos producidos por las fases anteriores y genera código de bajo nivel que puede ser ejecutado por la máquina. El generador de código suele generar código en lenguaje ensamblador o código máquina. En otros casos, genera bytecode, como en Java o cuando se utiliza el compilador just-in-time (JIT) de Python.

Finalmente, puede haber cero o más fases de **optimización del código (*code optimization*)**. Estas son, desde un punto de vista teórico, opcionales, pero suelen incluirse por defecto en los compiladores modernos. En esta fase, el compilador analiza el código generado e intenta mejorar su eficiencia aplicando diversas técnicas de optimización. Algunos ejemplos de optimizaciones son:

- plegado de constantes (*constant folding*) [Aho et al., 2014, Chap. 8.5.4],
- desenrollado de bucles (*loop unrolling*) [Aho et al., 2014, Chap. 10.5],
- asignación de registros (*register allocation*) [Aho et al., 2014, Chap. 8.1.4],
- propagación de constantes (*constant propagation*) [Aho et al., 2014, Chap. 9],
- análisis de vida útil (*livenes analysis*) [Aho et al., 2014, Chap. 9],
- y muchos más...

Las optimizaciones *locales* de código se refieren a mejoras dentro de un bloque básico, mientras que la optimización *global* de código es cuando las mejoras tienen en cuenta lo que ocurre más allá de un bloque básico. En Rust, un ejemplo de optimización global es la optimización del tiempo de enlace (link time optimization (LTO)) [Huss, 2020].

La Fig. 1.15 tomada de [Aho et al., 2014] visualiza intuitivamente las fases del compilador descritas en esta sección.

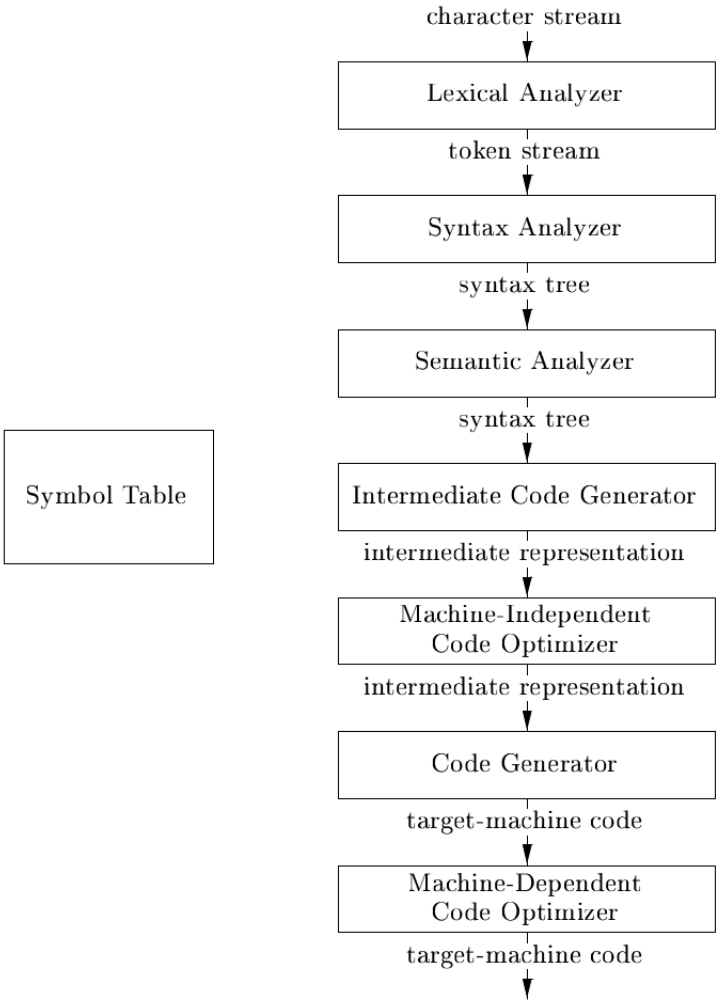

Figura 1.15: Fases de un compilador.

En la práctica, las fases pueden tener límites poco claros. Pueden solaparse y algunas pueden saltarse por completo. En secciones posteriores estudiaremos la arquitectura del compilador de Rust *rustc* y explicaremos su arquitectura en términos generales.

1.7. Verificación de modelos

La comprobación de modelos es una técnica utilizada en el desarrollo de software para verificar formalmente la corrección del comportamiento de un sistema con respecto a sus especificaciones

o requisitos. Consiste en construir un modelo matemático del sistema y analizarlo para garantizar que cumple ciertas propiedades, como la exclusión mutua al acceder a recursos compartidos, la ausencia de carreras de datos (*data races*) y la ausencia de deadlocks.

El proceso de comprobación de modelos comienza construyendo un modelo de estado finito del sistema, típicamente utilizando un lenguaje formal, en el caso de este trabajo el lenguaje de las redes de Petri. El modelo captura el comportamiento del sistema y las propiedades que deben verificarse. El siguiente paso es realizar una búsqueda exhaustiva del espacio de estados del modelo para asegurarse de que se han considerado todos los comportamientos posibles. Esta búsqueda puede realizarse automáticamente utilizando herramientas de software especializadas.

Durante la búsqueda, el verificador de modelos busca contraejemplos, es decir secuencias de eventos que violen las especificaciones del sistema. Si se encuentra un contraejemplo, el verificador de modelos proporciona información sobre el estado del sistema en el momento de la violación, lo que ayuda a los desarrolladores a identificar y solucionar el problema.

La comprobación de modelos se ha convertido en una técnica ampliamente aplicada en el desarrollo de sistemas de software críticos, como los dispositivos médicos, los sistemas financieros y de control aeroespacial [Carreño and Muñoz, 2005, Monzon and Fernandez-Sanchez, 2009] y de automóviles [Perronnet et al., 2019]. Al verificar la corrección del software antes de su despliegue, los desarrolladores pueden garantizar que el sistema cumple sus requisitos y es seguro de usar.

Una de las principales ventajas de la comprobación de modelos es que proporciona un enfoque formal y riguroso para verificar la correctitud del software. A diferencia de los métodos de prueba (test) tradicionales, que sólo pueden demostrar la presencia de errores, la comprobación de modelos puede demostrar la ausencia de errores. Esto es especialmente relevante para sistemas de seguridad crítica como los mencionados anteriormente en los que un solo error puede tener consecuencias catastróficas para vidas humanas. La comprobación de modelos también puede automatizarse, lo que permite a los desarrolladores verificar de forma rápida y eficaz la correctitud de sistemas de software complejos. Esto reduce el tiempo y el coste del desarrollo de software y aumenta la confianza en el sistema.

Se sabe que las herramientas formales de verificación de software se aplican actualmente en unos pocos campos muy específicos en los que se requiere una demostración formal de la correctitud del sistema. [Reid et al., 2020] habla de la importancia de acercar las herramientas de verificación a los desarrolladores mediante un enfoque que busque maximizar la relación coste- beneficio de su uso. Se presentan mejoras en la usabilidad de las herramientas existentes y enfoques para incorporar su uso a la rutina del desarrollador. El paper parte de la premisa de que, desde el punto de vista del desarrollador, la verificación puede verse como un tipo diferente de prueba unitaria o de integración. Por lo tanto, es de suma importancia que la ejecución de la verificación sea lo más sencilla posible y que se proporcione retroalimentación (feedback) al desarrollador con rapidez durante el proceso de desarrollo para aumentar su adopción.

La principal conclusión de esta sección es que la comprobación de modelos podría aportar

mejoras sustanciales en términos de mayor seguridad y fiabilidad de los sistemas de software. Estos objetivos se alinean con las metas del lenguaje de programación Rust y los objetivos de este trabajo. Detectar los bloqueos y las señales perdidas en el código fuente en tiempo de compilación podría ayudar a los desarrolladores a evitar errores difíciles de encontrar y a obtener feedback rápidamente sobre el uso correcto de las primitivas de sincronización, ahorrando tiempo y, además, dinero en el proceso de desarrollo. Un objetivo concreto de este trabajo es hacer que la herramienta sea fácil de usar y de empezar a utilizar, de modo que su adopción beneficie a la comunidad entera de desarrolladores de Rust.

Capítulo 2

Estado del arte

En este capítulo, se revisa brevemente la literatura sobre la verificación formal del código Rust y el modelado de redes de Petri para la detección de deadlocks. Algunas de estas publicaciones anteriores contienen enfoques que han guiado este trabajo.

En las dos secciones sucesivas examinaremos las herramientas existentes, su alcance y sus objetivos en comparación con la herramienta desarrollada en esta tesis.

A continuación, se ofrece un estudio de las bibliotecas de redes de Petri existentes en el ecosistema Rust a principios de 2023 para justificar la necesidad de implementar una biblioteca desde cero.

Posteriormente, exploramos la comunidad de investigadores detrás del Concurso de Verificación de Modelos (Model Checking Contest (MCC)) y los verificadores de modelos que participan en él para confirmar el potencial de estas herramientas para analizar modelos de redes de Petri de tamaño significativo. Esto es relevante ya que el verificador de modelos actúa como backend de la herramienta desarrollada en este trabajo.

Por último, se presentan tres de los formatos de archivo existentes para el intercambio de redes de Petri y se explica su finalidad en el contexto de este trabajo.

2.1. Verificación formal de código Rust

Existen numerosas herramientas de verificación automática disponibles para el código Rust. Una primera aproximación recomendable al tema es la encuesta elaborada por Alastair Reid, investigador de Intel. En ella se enumera explícitamente que la mayoría de las herramientas de verificación formal no soportan la concurrencia [Reid, 2021].

El intérprete *Miri*[1] desarrollado por el proyecto Rust en GitHub es un intérprete experimental

[1] https://github.com/rust-lang/miri

para la representación intermedia del lenguaje Rust ("Mid-level Intermediate Representation", comúnmente conocida como "MIR") que permite ejecutar binarios estándar de proyectos de *cargo* en una forma granularizada, instrucción por instrucción, para comprobar la ausencia de Comportamientos Indefinidos (Undefined Behavior (UB)) y otros errores en el manejo de la memoria. Detecta fugas de memoria, accesos a memoria no alineados, carreras de datos y violaciones de precondiciones o invariantes en el código marcado como inseguro (`unsafe`).

[Toman et al., 2015] presenta un verificador formal para Rust que no requiere modificaciones en el código fuente. Se probó en versiones anteriores de módulos de la biblioteca estándar de Rust. Como resultado se detectaron errores en el uso de la memoria en código Rust `unsafe` que, en circunstancias reales, el equipo de desarrollo tardó meses en descubrir manualmente. Esto ejemplifica la importancia de utilizar herramientas de verificación automática para complementar las revisiones manuales del código (*code reviews*).

[Kani Project, 2023] es otra herramienta popular para la verificación formal de código Rust destinada a comprobar los bloques inseguros (`unsafe`) a nivel de bits. Ofrece un framework de pruebas análogo al framework de pruebas proporcionado por Rust. Además, dispone de un plugin para *cargo* y VS Code.

Como explica la documentación del repositorio[2], Kani verifica (entre otros):

- Uso seguro de la memoria, por ejemplo, desreferencias de punteros nulos (*Memory safety, e.g., null pointer dereferences*)
- Aserciones especificadas por el usuario (*User-specified assertions, i.e.,* `assert!(...)`)
- La ausencia de panics, por ejemplo, `unwrap()` en valores `None`
- La ausencia de algunos tipos de comportamiento inesperado, por ejemplo, desbordamientos aritméticos (*arithmetic overflows*)

Sin embargo, los programas concurrentes están actualmente fuera de alcance[3]. La conclusión es que Kani ofrece una CLI fácil de usar y un framework de pruebas que se integran perfectamente en el proceso de desarrollo. Sirve como ilustración de las capacidades de la comprobación de modelos en el desarrollo de software moderno.

2.2. Detección de deadlocks mediante redes de Petri

La prevención de los bloqueos es una de las estrategias clásicas para abordar este problema fundamental en la programación concurrente, como se explica en la Sec. 1.4.2. El principal problema del enfoque de detectar los bloqueos antes de que se produzcan es probar que se detecta el tipo de bloqueo deseado en todos los casos y que no se producen falsos negativos en el proceso.

[2] https://github.com/model-checking/kani
[3] https://model-checking.github.io/kani/rust-feature-support.html

El enfoque basado en redes de Petri, al ser un método formal, satisface estas condiciones. Sin embargo, la dificultad de su adopción radica principalmente en la practicabilidad de la solución debido al gran número de estados posibles en un proyecto de software real.

En [Karatkevich and Grobelna, 2014], se propone un método para reducir el número de estados explorados durante la detección de bloqueos mediante el análisis de alcanzabilidad. Esta heurística ayuda a mejorar el rendimiento del enfoque basado en redes de Petri. Otra optimización se presenta en [Küngas, 2005]. El autor propone un método de orden polinómico muy prometedor para evitar el problema de la explosión de estados que subyace al algoritmo ingenuo de detección de deadlocks. A través de un algoritmo que abstrae una red de Petri dada a una representación más simple, se obtiene una jerarquía de redes de tamaño creciente para las que la verificación de la ausencia de bloqueos es sustancialmente más rápida. Se trata, dicho crudamente, de una estrategia de "divide y vencerás" que comprueba la ausencia de deadlocks en partes de la red para construir después la verificación del conjunto final añadiendo partes a la pequeña red inicial.

A pesar de las advertencias mencionadas anteriormente, el uso de las redes de Petri como método formal de verificación de software se ha establecido desde finales de la década de 1980. Las redes de Petri permiten un modelado intuitivo de las primitivas de sincronización, como el envío de un mensaje o la espera de la recepción de un mensaje. En [Heiner, 1992] encontrará ejemplos de estas sencillas redes con un comportamiento correspondientemente simple. Estas redes son bloques de construcción que pueden combinarse para formar un sistema más complejo.

Para poner en práctica estos modelos, existen dos posibilidades:

- Una es diseñar el sistema en términos de redes de Petri y luego traducir las redes de Petri al código fuente.
- La otra consiste en traducir el código fuente existente a una representación de red de Petri y, a continuación, verificar que el modelo de red de Petri satisface las propiedades deseadas.

A efectos de este trabajo, nos interesa esta última. Este enfoque no es novedoso. Ya se ha implementado para otros lenguajes de programación como C y Rust, como se ve en la bibliografía existente.

En [Kavi et al., 2002] y [Moshtaghi, 2001], se describe una traducción de algunas primitivas de sincronización disponibles como parte de la biblioteca POSIX de hilos (`pthread`) en C a redes de Petri. En concreto, la traducción admite:

- La creación de hilos con la función `pthread_create` y el manejo de la variable de tipo `pthread_t`.
- La operación de unión de hilos con la función `pthread_join`.
- La operación de adquirir un mutex con `pthread_mutex_lock` y su eventual liberación manual con `pthread_mutex_unlock`.

- Las funciones `pthread_cond_wait` y `pthread_cond_signal` para trabajar con condition variables.

Lamentablemente, el código fuente de esta biblioteca llamada "C2Petri" no se encuentra en línea, ya que las publicaciones son bastante antiguas.

En una tesis de máster más reciente, [Meyer, 2020] establece las bases de una semántica de redes de Petri para el lenguaje de programación Rust. Sin embargo, centra sus esfuerzos en el código de un solo hilo, limitándose a la detección de los deadlocks causados por la ejecución de la operación de `lock` dos veces sobre el mismo mutex en el hilo principal. Desafortunadamente, el código disponible en GitHub[4] como parte de la tesis ya no es válido para la nueva versión de *rustc*, puesto que las partes internas del compilador han cambiado significativamente en los últimos tres años.

En un *pre-print* de finales de 2022, [Zhang and Liua, 2022] implementan una traducción del código fuente de Rust a redes de Petri para encontrar deadlocks. La traducción se centra en los bloqueos muertos causados por dos tipos de bloqueos de la biblioteca estándar: `std::sync::Mutex` y `std::sync::RwLock`. La red de Petri resultante se expresa en el lenguaje de marcado de redes de Petri (Petri Net Markup Language (PNML)) y se introduce en el verificador de modelos Platform Independent Petri net Editor 2 (PIPE2)[5] para realizar el análisis de alcanzabilidad. Las llamadas a funciones se manejan de una forma muy diferente a la de este trabajo y las señales perdidas no se modelan en absoluto. El código fuente de su herramienta, denominada TRustPN, no está disponible públicamente en el momento de escribir este artículo. A pesar de estas limitaciones, los autores ofrecen un estudio muy detallado y actualizado de las herramientas de análisis estático para la verificación de código Rust que podría resultar atractivo para el lector interesado en ahondar en esta temática. Además, enumeran varios trabajos dedicados a formalizar la semántica del lenguaje de programación Rust que quedan fuera del alcance de este trabajo.

2.3. Bibliotecas de redes de Petri en Rust

Como parte del desarrollo de la traducción del código fuente a una red de Petri, es necesario utilizar una biblioteca de redes de Petri para el lenguaje de programación Rust. Una búsqueda rápida de los paquetes disponibles en *crates.io*[6], GitHub y GitLab reveló que, por desgracia, no existe ninguna biblioteca bien mantenida.

Se encontraron algunos simuladores de redes de Petri como:

- pns[7]: Programado en C. No ofrece la opción de exportar la red resultante a un formato

[4] https://github.com/Skasselbard/Granite
[5] https://pipe2.sourceforge.net/
[6] https://crates.io/
[7] https://gitlab.com/porky11/pns

estándar.

- PetriSim[8]: Un antiguo simulador DOS/PC programado en Bordland Pascal.
- WOLFGANG[9]: Un editor de redes de Petri en Java, mantenido por el Departamento de Informática de la Universidad de Friburgo, Alemania.

Desafortunadamente ninguno de ellos cumple los requisitos de la tarea.

Considerando que una red de Petri es un grafo, se evaluó la posibilidad de utilizar una biblioteca de grafos y modificarla para adaptarla a los objetivos de este trabajo. Se encontraron dos bibliotecas de grafos en Rust:

- petgraph[10]: La biblioteca más utilizada para gráficos en *crates.io*. Ofrece una opción para exportar al formato DOT.
- gamma[11]: Inestable y sin cambios desde 2021. No ofrece la posibilidad de exportar el gráfico.

Ninguna de las posibilidades satisface el requisito de exportar la red resultante al formato PNML. Además, si se utiliza una biblioteca de grafos, las operaciones de una red de Petri deben implementarse como un *wrapper* alrededor de un grafo, lo que reduce la posibilidad de optimizaciones para nuestro caso de uso y dificulta la extensibilidad a largo plazo del proyecto.

En conclusión, es imperativo implementar una biblioteca de redes de Petri en Rust desde cero como un proyecto independiente. Esto aporta una herramienta más a la comunidad que podría reutilizarse en el futuro.

2.4. Verificadores de modelos

La elección de un verificador de modelos adecuado es una parte vital de este trabajo porque es el responsable de verificar la ausencia de bloqueos. Afortunadamente se han desarrollado varios comprobadores de modelos para analizar redes de Petri.

El Model Checking Contest (MCC) [Kordon et al., 2021] organizado en la Universidad de la Sorbona de París es una gran fuente de verificadores de modelos de última generación. Se trata de un concurso anual en el que los verificadores de modelos presentados se ejecutan sobre una serie de modelos de redes de Petri procedentes del mundo académico y de la industria[12]. Estos modelos han sido aportados por muchas personas a lo largo de un periodo de más de una década

[8] https://staff.um.edu.mt/jskl1/petrisim/index.html
[9] https://github.com/iig-uni-freiburg/WOLFGANG
[10] https://docs.rs/petgraph/latest/petgraph/
[11] https://github.com/metamolecular/gamma
[12] https://mcc.lip6.fr/2023/models.php

y el número total de puntos de referencia ha crecido paulatinamente a medida que se han ido añadiendo nuevos modelos.

Cada año, los puntos de referencia incluyen redes de lugares/transiciones (place/transition nets (P/T nets)), es decir, redes de Petri, y redes de Petri coloreadas (Colored Petri nets (CPN)). El número de lugares en las redes puede oscilar entre una docena y más de 70000 y las transiciones entre menos de un centenar y más de un millón. Esto pone de manifiesto la aplicabilidad práctica de los verificadores de modelos que participan en el concurso.

Los resultados se publican en la página web oficial (véase por ejemplo [Kordon et al., 2022]) y consisten en:

1. una lista de las herramientas cualificadas que participaron,
2. las técnicas aplicadas en cada una de las herramientas,
3. una sección dedicada a detallar las condiciones experimentales en las que se desarrolló el concurso (el hardware utilizado y el tiempo necesario para completar las ejecuciones),
4. los resultados en forma de tablas, gráficos e incluso los registros de ejecución de cada programa,
5. la lista de ganadores de cada categoría,
6. un análisis de la fiabilidad de las herramientas basado en la comparación de los resultados.

Un breve vistazo a las diapositivas de la edición de 2022[13] reproducidas en la Fig. 2.1 ilustra que varios verificadores de modelos han demostrado una participación ininterrumpida, con ejemplos notables que incluyen:

- Tool for Verification of Timed-Arc Petri Nets (TAPAAL) mantenida por la Universidad de Aalborg en Dinamarca[14], ganadora de una medalla de oro en la edición de 2023.
- Low-Level Petri Net Analyzer (LoLA) mantenido por la Universidad de Rostock en Alemania[15], ganador en ediciones anteriores y fue utilizado base para otros verificadores de modelos.
- ITS-tools [Thierry Mieg, 2015], que también se combinó con LoLA y obtuvo medallas en 2020[16].

Estas observaciones indican colectivamente la madurez y vitalidad de la comunidad de verificadores de modelos. El establecimiento de un panorama de herramientas bien desarrollado, fomentado por la colaboración y la difusión de código abierto de resultados, *benchmarks* y técnicas, presenta una valiosa oportunidad para aprovechar estas herramientas en el ámbito del

[13] https://mcc.lip6.fr/2022/pdf/MCC-PN2022.pdf
[14] https://www.tapaal.net/
[15] https://theo.informatik.uni-rostock.de/theo-forschung/tools/lola/
[16] https://github.com/yanntm/its-lola

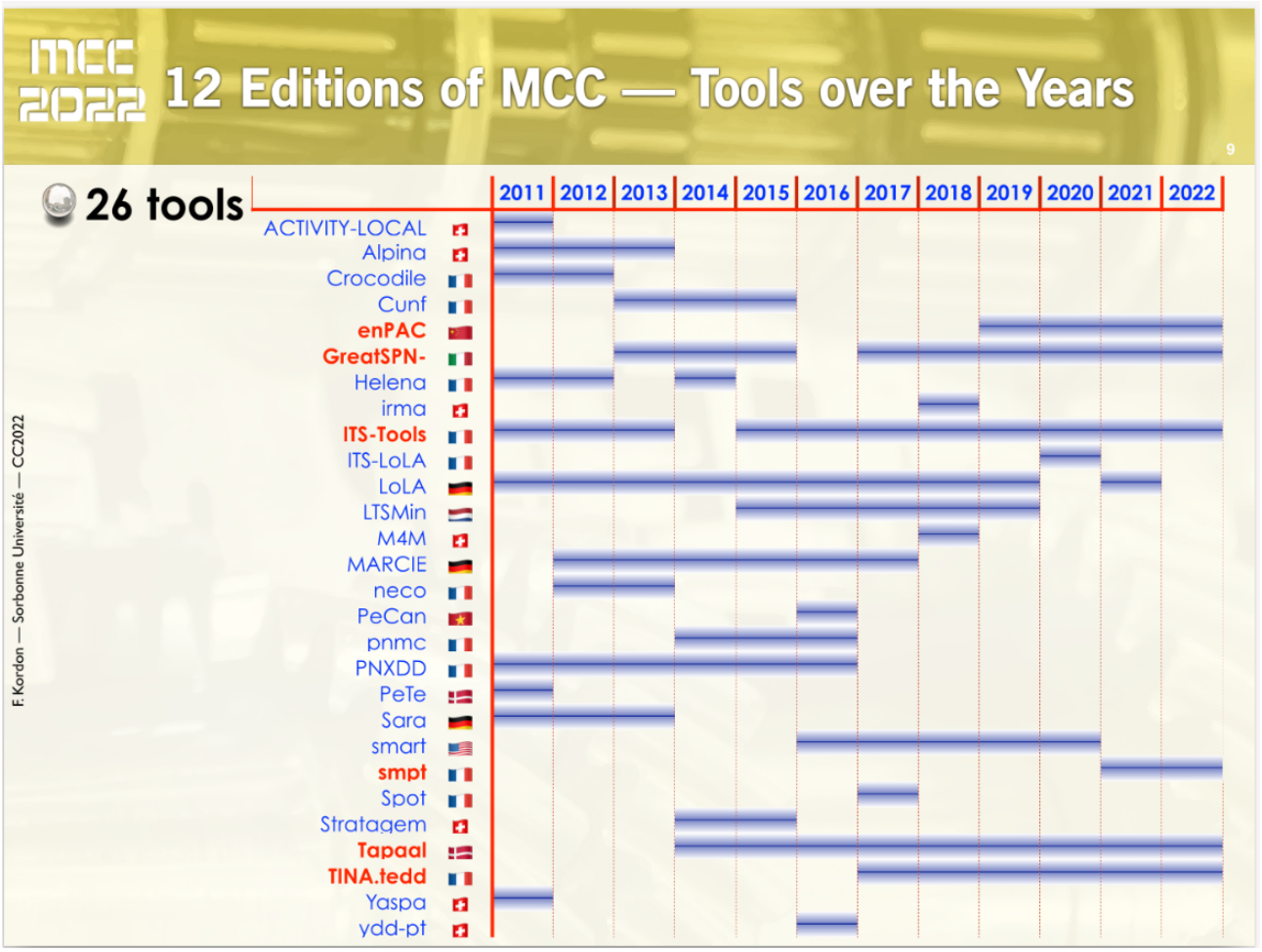

Figura 2.1: Participación de los verificadores de modelos en el MCC a lo largo de los años.

desarrollo de software. Concretamente, en el contexto de integrarlas como backends para un traductor para un lenguaje de programación específico que se encargue de automatizar el proceso de creación de modelos de redes de Petri. Capitalizando los esfuerzos académicos invertidos en los verificadores de modelos, se puede lograr una mayor seguridad y fiabilidad en los proyectos de software.

2.5. Formatos de archivo para intercambio de redes de Petri

Como se ha observado en el capítulo anterior, las redes de Petri son una herramienta muy utilizada para modelar sistemas de software. Sin embargo, debido a las diferentes clases de redes de Petri (redes de Petri simples, redes de Petri de alto nivel, redes de Petri con tiempo,

redes de Petri estocásticas, redes de Petri coloreadas, por nombrar algunas), diseñar un formato de archivo de intercambio estandarizado compatible con todas las aplicaciones ha resultado todo un reto. Una de las razones es que las redes de Petri pueden implementarse y representarse de múltiples formas, en función de los objetivos específicos, visto que son un tipo de grafo.

Para garantizar un cierto grado de interoperabilidad entre la herramienta desarrollada en el marco de esta tesis y otras herramientas existentes y futuras, es primordial investigar qué formatos de archivo sería más conveniente soportar. El objetivo es admitir formatos de archivo que sean adecuados tanto para el análisis como para la visualización, permitiendo la posibilidad de ampliación a formatos adicionales en el futuro, a través de una API bien definida en la biblioteca de redes de Petri. Una revisión de la literatura condujo a tres formatos de archivo relevantes que se presentan a continuación.

2.5.1. Petri Net Markup Language

El Petri Net Markup Language (PNML)[17] es un formato de archivo estándar diseñado para el intercambio de redes de Petri entre distintas herramientas y aplicaciones de software. Su desarrollo se inició en el "Meeting on XML/SGML based Interchange Formats for Petri Nets" celebrada en Aarhus en junio de 2000 [Jüngel et al., 2000, Weber and Kindler, 2003] con el objetivo de proporcionar un formato estandarizado y ampliamente aceptado para redes de Petri. PNML es una norma ISO que consta, a partir de 2023, de tres partes:

- ISO/IEC 15909-1:2004[18] (y su última revisión ISO/IEC 15909-1:2019[19]) para conceptos, definiciones y notación gráfica.
- ISO/IEC 15909-2:2011[20] para la definición de un formato de transferencia basado en XML.
- ISO/IEC 15909-3:2021[21] para las extensiones y los mecanismos de estructuración.

Se ha convertido en un estándar *de facto* para intercambiar modelos en redes de Petri entre diferentes herramientas y sistemas. Es el resultado de muchos años de duro trabajo para unificar la notación, tal y como se expone en [Hillah and Petrucci, 2010].

PNML ha sido diseñado para ser un formato flexible y extensible que pueda representar diferentes clases de redes de Petri, incluidas las redes de Petri simples y las redes de Petri de alto nivel. Se basa en Extensible Markup Language (XML), lo que facilita su lectura y análisis tanto por humanos como por máquinas. Además, PNML admite el uso de metadatos para proporcionar información adicional sobre los modelos de redes de Petri, como la autoría, la fecha de creación e información sobre licencias.

[17] https://www.pnml.org/
[18] https://www.iso.org/standard/38225.html
[19] https://www.iso.org/standard/67235.html
[20] https://www.iso.org/standard/43538.html
[21] https://www.iso.org/standard/81504.html

has been designed to be a flexible and extensible format that can represent different classes of Petri nets, including simple Petri nets and high-level Petri nets. It is based on the which makes it easy to read and parse by humans and machines alike. Additionally, supports the use of metadata to provide additional information about the Petri net models, such as authorship, date of creation, and licensing information.

El desarrollo de PNML ha mejorado significativamente la interoperabilidad y el intercambio de modelos de redes de Petri entre diferentes herramientas y sistemas. Antes de la adopción de PNML el intercambio de modelos de redes de Petri era una tarea ardua puesto que las distintas herramientas utilizaban formatos propietarios que a menudo eran incompatibles entre sí. El PNML ha simplificado enormemente este proceso, permitiendo a investigadores y profesionales compartir y colaborar en modelos de redes de Petri con facilidad. Su uso también ha facilitado el desarrollo de nuevas herramientas y aplicaciones de software para redes de Petri porque proporciona un formato estándar que puede ser analizado y procesado fácilmente por distintos sistemas. En particular es el formato utilizado en [Zhang and Liua, 2022] y está soportado en [Meyer, 2020].

2.5.2. Formato GraphViz DOT

El formato DOT es un lenguaje de descripción de grafos utilizado para crear representaciones visuales de grafos y redes, que forma parte de la suite de código abierto GraphViz[22]. Fue creado a principios de la década de 1990 en AT&T Labs Research como un lenguaje sencillo, conciso y legible por humanos para la descripción de grafos. La suite GraphViz proporciona varias herramientas para trabajar con archivos DOT, incluida la capacidad de generar automáticamente diseños para gráficos complejos y de exportar visualizaciones en diversos formatos, como PNG, PDF y SVG.

DOT puede utilizarse para representar redes de Petri en un formato gráfico, lo que facilita la visualización de la estructura y el comportamiento del sistema que se está modelando. Resulta especialmente útil para visualizar redes de Petri de gran tamaño, ya que el usuario puede navegar por la imagen para comprender cómo fluyen las marcas por la red.

El formato DOT está basado en texto plano y es fácil de usar, convirtiéndolo en una opción popular para generar representaciones visuales de gráficos. Esta facilidad también significa que los archivos DOT pueden ser generados fácilmente por programas y pueden ser leídos por una amplia gama de herramientas de software, un aspecto esencial para la interoperabilidad. Además, DOT permite especificar diversas propiedades de los grafos, como formas de nodos, colores y estilos [Gansner et al., 2015] que pueden utilizarse para representar diferentes aspectos de una red de Petri, como lugares, transiciones y arcos. Esta flexibilidad a la hora de especificar las propiedades visuales también permite a los usuarios personalizar la visualización según sus necesidades y resaltar características particulares de la red de Petri que sean relevantes para su análisis.

[22] https://graphviz.org/

2.5.3. LoLA - Low-Level Petri Net Analyzer

Low-Level Petri Net Analyzer (LoLA) [Schmidt, 2000] es un verificador de modelos de última generación cuyo desarrollo comenzó en 1998 en la Universidad Humboldt de Berlín. Actualmente lo mantiene la Universidad de Rostock y se publica bajo la Licencia Pública General Affero de GNU. LoLA es una herramienta que puede comprobar si un sistema satisface una propiedad dada expresada en Computational Tree Logic* (CTL*). Su punto fuerte es la evaluación de propiedades sencillas como la libertad de bloqueo (*deadlock freedom*) o la alcanzabilidad, tal y como se indica en la página web. Este es el verificador de modelos utilizado en [Meyer, 2020] y en este trabajo. En consecuencia, es necesario implementar el formato de archivo requerido por la herramienta. En la Sec. 5.2 se presentan ejemplos.

Capítulo 3

Diseño de la solución propuesta

Una vez cubiertos los temas de fondo pertinentes, podemos proceder a profundizar en los aspectos específicos del diseño del proceso de traducción. El diseño está marcado por tres opciones arquitecturales cruciales sobre las que se profundizará en este capítulo:

1. La decisión de utilizar el compilador Rust como backend para la traducción.
2. Basar la traducción en el Mid-level Intermediate Representation (MIR).
3. Hacer un *inlining* de las llamadas a funciones en la red de Petri.

A lo largo de este capítulo, analizaremos en profundidad los mecanismos internos del compilador de Rust y sus etapas de compilación relevantes para este trabajo.

3.1. En busca de un backend

Para ponerlo de forma sucinta, existen dos enfoques para traducir código Rust a redes de Petri. La primera opción es crear un traductor desde cero, mientras que la segunda consiste en basarse en una herramienta ya existente.

La primera opción puede parecer atractiva al principio, teniendo en cuenta que da al desarrollador la libertad de moldear la herramienta según sus deseos. Se pueden añadir funciones según las necesidades y adaptar las estructuras de datos al propósito específico. Sin embargo, esta flexibilidad tiene un alto precio. Para dar soporte a un subconjunto razonable del lenguaje de programación Rust, es necesario invertir grandes cantidades de esfuerzo en la tarea. Las construcciones complejas del lenguaje, como las macros, los *generics* o mismo el rico sistema de tipos, deben ser comprendidas en sus detalles más intrincados para poder ser traducidas con eficacia. El resultado es, esencialmente, un nuevo compilador para código Rust. Teniendo en cuenta que el compilador de Rust se desarrolló a lo largo de muchos años y con el apoyo de una gran comunidad de colaboradores, queda claro que este camino no es más que una duplicación

de trabajo. De hecho, es una labor hercúlea que requeriría la dedicación a tiempo completo de un equipo completo para mantener al día de los cambios más recientes en el lenguaje Rust y en el compilador.

Por otro lado, existe la posibilidad de integrarse con el compilador de Rust existente, que está disponible bajo una licencia de código abierto y su documentación es extensa y se actualiza con regularidad. Esto libera en parte a la implementación de tener que ocuparse de los cambios en el lenguaje, lo que da más tiempo para centrarse en las características que añaden valor a los usuarios. De ahí que el compilador desempeñe el papel de *backend* en el que se apoya el análisis estático. Por supuesto, esto requiere aprender las interioridades del compilador, pero no es la primera vez que una herramienta se propone ello. A modo de ejemplo, el linter oficial de Rust, *clippy*[1], analiza el código Rust en busca de construcciones incorrectas, ineficaces o no idiomáticas. Se trata de una herramienta muy valiosa para los desarrolladores que va más allá de las comprobaciones estándar realizadas durante la compilación.

Dar soporte a todas las características del lenguaje desde el principio y colaborar con la comunidad es clave para el éxito de la solución propuesta. Por lo tanto, es aconsejable integrarse con el ecosistema existente y reutilizar todo el trabajo posible. Por todo ello, este proyecto se basa en *rustc*. A continuación estudiaremos con más detalle los componentes centrales del compilador de Rust.

3.2. El compilador de Rust: *rustc*

El compilador de Rust, *rustc*, se encarga de traducir el código Rust en código ejecutable. Sin embargo, *rustc* no es un compilador tradicional en el sentido de que realiza múltiples pasadas sobre el código, como se describe en la Sec. 1.6. En su lugar, *rustc* está construido sobre un sistema basado en consultas que soporta la compilación incremental.

En el sistema de consulta de *rustc*, el compilador calcula un grafo de dependencias entre los artefactos de código, incluidos los archivos fuente, los crates y los artefactos intermedios, como los archivos objeto. A continuación, el sistema de consulta utiliza este grafo para recompilar eficientemente sólo aquellos artefactos que hayan cambiado desde la última compilación[2]. Esta compilación incremental puede reducir significativamente el tiempo de compilación de grandes proyectos, facilitando el desarrollo y la iteración del código Rust.

El sistema de consulta también permite al compilador de Rust realizar otras optimizaciones, como la memoización y el almacenamiento en caché de resultados intermedios. Por ejemplo, si el valor de retorno de una función se ha calculado antes, el sistema de consulta puede devolver el resultado almacenado en caché en lugar de volver a calcularlo, lo que reduce aún más el tiempo de compilación.

[1] https://github.com/rust-lang/rust-clippy
[2] https://rustc-dev-guide.rust-lang.org/queries/incremental-compilation.html

Otra elección de diseño importante en *rustc* es el *interning*. *Interning* es una técnica para almacenar cadenas de texto y otras estructuras de datos de forma eficiente en memoria. En lugar de almacenar varias copias de la misma cadena o estructura de datos, el compilador de Rust almacena sólo una copia en un *allocator* especial llamado *arena*. Las referencias a los valores almacenados en la arena se pasan de una parte a otra del compilador y pueden compararse de forma barata comparando punteros. Esto permite reducir el uso de memoria y acelerar las operaciones que comparan o manipulan cadenas de texto y estructuras de datos.

rustc utiliza la infraestructura del compilador LLVM[3] para realizar la generación de código de bajo nivel y la optimización. LLVM proporciona un marco flexible para compilar código a una variedad de *targets*, incluyendo código máquina nativo y WebAssembly (WASM). El compilador de Rust utiliza LLVM para optimizar el código en términos de rendimiento y generar código de alta calidad para una gran variedad de plataformas. En lugar de generar código máquina, sólo necesita generar la intermediate representation (IR) de LLVM del código fuente y luego ordenar a LLVM que lo transforme al objetivo de compilación (*compilation target*), aplicando las optimizaciones deseadas.

rustc está programado en Rust. Para compilar la versión más reciente del compilador y la versión más reciente de la biblioteca estándar que lo acompaña, se utiliza una versión ligeramente más antigua de *rustc* y de la biblioteca estándar. Este proceso se denomina *bootstrapping* e implica que uno de los principales usuarios de Rust es el propio compilador de Rust. Teniendo en cuenta que cada seis semanas se publica una nueva versión estable, el bootstrapping implica una gran complejidad y se describe detalladamente en la documentación[4] y en conferencias [Nelson, 2022] y tutoriales [Klock, 2022] hechos por miembros del Rust team.

3.2.1. Etapas de compilación

La existencia del sistema de consulta no implica que *rustc* no tenga fases de compilación en absoluto. Al contrario, se requieren varias etapas de compilación para transformar el código fuente de Rust en código máquina que pueda ejecutarse en una computadora. Estas etapas implican múltiples representaciones intermedias del programa, cada una optimizada para un propósito específico. A continuación describiremos brevemente estas etapas. Encontrará una descripción más completa en la documentación[5].

Lexado y análisis sintáctico

En primer lugar, el texto fuente en bruto de Rust es analizado por un *lexer* de bajo nivel. En esta etapa, el texto fuente se convierte en un flujo (*stream*) de unidades atómicas de código fuente conocidas como tokens.

[3] https://llvm.org/
[4] https://rustc-dev-guide.rust-lang.org/building/bootstrapping.html
[5] https://rustc-dev-guide.rust-lang.org/overview.html

A continuación se realiza el análisis sintáctico (*parsing*). El flujo de tokens se convierte en un AST. Aquí se produce el interning de los valores de cadena. La expansión de macros, la validación del AST, la resolución de nombres y el linting temprano también tienen lugar durante esta e t a p a . La representación intermedia resultante de esta etapa es, a fin de cuentas, el AST.

HIR lowering

Posteriormente, el AST se convierte en High-Level Intermediate Representation (HIR). Este proceso se conoce como "rebajar" (*lowering*). Esta representación se parece al código Rust pero con construcciones complejas convertidas en versiones más simples. Por ejemplo, todos los bucles `while` y `for` se convierten en versiones más simples con bucles `loop`.

La HIR se utiliza para realizar algunos pasos importantes:

1. *Inferencia de tipo (type inference)*: La detección automática del tipo de una expresión, por ejemplo, al declarar variables con `let`.
2. *Resolución de traits (trait solving)*: Garantizar que cada bloque de implementación (`impl`) hace referencia a un trait válido y existente.
3. *Comprobación de tipos (type checking)*: Este proceso convierte los tipos escritos por el usuario en la representación interna utilizada por el compilador. Es, en otras palabras, donde se internan los tipos. Después, utilizando esta información, se verifica la seguridad de tipos, la correctitud y la coherencia.

MIR lowering

En esta etapa, la HIR se rebaja a Mid-level Intermediate Representation (MIR), que se utiliza para el *borrow checking*. Como parte del proceso, se construye la Typed High-Level Intermediate Representation (THIR), que es una representación más fácil de convertir a MIR que la HIR.

La THIR es una versión aún más "desugarizada" (*desugared*) del HIR. Se utiliza para el *pattern matching* y el *exhaustiveness matching*. Es similar a la HIR pero con todos los tipos y llamadas a métodos explícitos. Además se incluyen desreferencias implícitas cuando es necesario.

Muchas optimizaciones se realizan sobre la MIR puesto que sigue siendo una representación muy genérica. Las optimizaciones son en algunos casos más fáciles de realizar sobre la MIR que sobre la posterior IR de LLVM.

Generación de código

Esta es la última etapa en la producción de un binario. Incluye la llamada a LLVM para la generación de código y las optimizaciones correspondientes. Para ello, la MIR se convierte en LLVM IR.

LLVM IR es la forma estándar de entrada para el compilador LLVM que utilizan todos los compiladores que utilizan LLVM, como el compilador de C *clang*. Es un tipo de lenguaje ensamblador bien anotado y diseñado para que otros compiladores puedan producirlo fácilmente. Además, está diseñado para ser lo suficientemente rico como para permitir a LLVM realizar varias optimizaciones sobre él.

LLVM transforma el LLVM IR a código máquina y aplica muchas más optimizaciones. Por último, los archivos objeto que contienen código ensamblador pueden enlazarse (*linking*) entre sí para formar el binario.

3.2.2. Rust nightly

Comprender el modelo de lanzamiento de versiones de Rust es indispensable para implementar con éxito la herramienta propuesta en este trabajo. La razón es que para utilizar las crates de *rustc* como dependencia en nuestro proyecto, debe compilarse con la versión *nightly*.

El compilador nightly de Rust se refiere a una compilación específica de *rustc* que se actualiza cada noche con los últimos cambios y mejoras pero que también incluye características experimentales o inestables que aún no forman parte de la versión estable. En Rust, el lenguaje y su biblioteca estándar se versionan utilizando un modelo de "tren de versiones" (*release train*), en el que existen tres canales de versiones principales: estable, beta y nightly[6].

La versión estable del compilador de Rust es la más utilizada y recomendada para su uso en producción. Pasa por un riguroso proceso de pruebas y estabilización para garantizar que proporciona una experiencia estable y fiable a los desarrolladores. La versión estable sólo incluye características y mejoras que han sido revisadas a fondo, testeadas y consideradas lo suficientemente estables para su uso en producción.

Por otro lado, el compilador nightly de Rust es la versión más *bleeding-edge*, en la que se introducen a diario nuevas características, correcciones de errores y cambios experimentales. Es utilizado por los desarrolladores y colaboradores del lenguaje Rust con fines de prueba y desarrollo, pero no se recomienda su uso en producción debido a la inestabilidad potencial y a la falta de soporte a largo plazo.

Cada característica exclusiva de la versión nightly está detrás de una *feature flag*. Sólo pueden utilizarse al compilar con la *toolchain* nightly. Las feature flags pueden habilitar

[6] https://forge.rust-lang.org/

- construcciones sintácticas que no están disponibles en la versión estable,
- funciones de biblioteca exclusivas de la versión nocturna,
- soporte para instrucciones de hardware específicas de un ISA o plataforma determinados,
- flags adicionales del compilador.

La lista completa de banderas de características se encuentra en [Rust Project, 2023e] y contiene más de 500 entradas en total. De forma más concisa, el lenguaje Rust utilizado dentro de *rustc* es un superconjunto del lenguaje Rust estable utilizado fuera de él. Estas diferencias deben tenerse en cuenta cuando se trabaje en el compilador o se construya software que dependa directamente del compilador.

3.3. Selección de un punto de partida adecuado para la traducción

En esta sección, se elucidan los motivos para seleccionar la Mid-level Intermediate Representation (MIR) como punto de partida para la traducción a una red de Petri. Esta elección de diseño arquitectural se justifica por varias razones.

3.3.1. Beneficios

En primer lugar, la MIR es la IR más baja en rustc que aún es independiente de la arquitectura de la computadora. Captura la semántica del código Rust después de que haya sido sometido a una serie de pases de optimización sin depender de los detalles de cualquier máquina en particular. Al interceptar la traducción en esta fase, la herramienta de análisis estático aprovecha las ventajas de estas optimizaciones, como el plegado de constantes (*constant folding*), la eliminación de código muerto y el *inlining*, lo que da como resultado una representación de la red de Petri más eficiente y, en general, más pequeña.

En segundo lugar, interceptar la compilación una vez completadas las etapas anteriores ofrece una ventaja en términos de eficiencia y reutilización del código. En esta etapa, el compilador de Rust ya ha realizado pasos cruciales como el *borrow checking*, la comprobación de tipos, la monomorfización del código genérico y la expansión de macros, entre otros. Estos pasos consumen muchos recursos e implican un análisis complejo del código Rust para garantizar su correctitud y seguridad. Reimplementar estos pasos en nuestra herramienta desde cero sería redundante y llevaría mucho tiempo. Requeriría duplicar los esfuerzos del compilador de Rust y podría introducir posibles incoherencias o errores. Al construir sobre el MIR existente, aprovechamos al máximo el trabajo ya realizado por *rustc*. Esto no sólo ahorra esfuerzo, sino que también alinea nuestra herramienta de análisis estático con el mismo nivel de correctitud y seguridad que el compilador de Rust.

En tercer lugar, simplifica la tarea de mantenimiento de mantenerse al día con las continuas incorporaciones al lenguaje Rust y a su compilador. Rust es un lenguaje en rápida evolución y su compilador se actualiza constantemente con nuevas características, correcciones de errores y mejores de performance. Reutilizar la MIR significa que nuestra herramienta puede beneficiarse de estas actualizaciones sin tener que implementar y mantener esos cambios de forma independiente. Esto proporciona en general una solución de análisis estático más robusta y fiable.

Adicionalmente, como se explicará en la siguiente sección, la MIR se basa en el concepto de grafo de flujo de control (control flow graph (CFG)), o en otras palabras, un tipo de grafo que se encuentra en los compiladores. Esto significa que tanto la MIR como las redes de Petri son representaciones gráficas, lo que hace que la MIR sea especialmente adecuada para una traducción. Tanto la MIR como las redes de Petri pueden considerarse modelos gráficos que capturan las relaciones e interacciones entre diferentes entidades. El grafo MIR representa el flujo de ejecución subyacente dentro de un programa Rust, mientras que una red Petri captura las transiciones de estado y las ocurrencias de eventos en un sistema. Consecuentemente, resulta más fácil convertir la MIR en una red de Petri, dado que la estructura del grafo y las relaciones ya están presentes. Esto permite un proceso de traducción más directo y eficiente sin tener que crear una estructura de grafos de la nada, lo que resulta en una mejor integración entre el MIR y el modelo de red de Petri para la detección de deadlocks.

Para concluir, trabajar con la MIR crea sinergias con la compilación incremental y el análisis modular. De hecho, una de las razones por las que se introdujo MIR en primer lugar fue la compilación incremental [Matsakis, 2016]. Aunque no es obligatorio en la implementación inicial, la herramienta podría beneficiarse de la compilación incremental y realizar análisis por crate/por módulo, lo que permitiría un análisis más rápido y eficiente de grandes bases de código Rust.

3.3.2. Limitaciones

A pesar de los numerosos beneficios, el enfoque de basar la traducción en la Mid-level Intermediate Representation (MIR) tiene algunas limitaciones.

La más importante es que la MIR está sujeta a cambios. No se ofrecen garantías de estabilidad en cuanto a cómo se traducirá el código Rust a MIR o cuáles son sus elementos constitutivos. Se trata de detalles internos que los desarrolladores del compilador se reservan para sí mismos. En resumen, la MIR como interfaz no es estable. A medida que se sigue trabajando en el compilador, la MIR sufre modificaciones para incorporar nuevas características del lenguaje, optimizaciones o correcciones de errores, lo que puede requerir frecuentes actualizaciones y ajustes en el proceso de traducción, aumentando el coste de mantenimiento.

En el transcurso de este proyecto esta situación se produjo en varias ocasiones. A modo de ejemplo, en el periodo comprendido entre mediados de febrero de 2023 y mediados de abril de

2023, el código se modificó 7 veces para dar cabida a estos cambios. Siempre fueron de unas pocas líneas de código y se detectaron mediante pruebas. Hablaremos de cómo las pruebas desempeñan un papel importante a la hora de lidiar con estos cambios en la Sec. 5.2.

En la misma línea, [Meyer, 2020] también se basó en la MIR pero no incorporó pruebas para hacer frente a las versiones nightly más recientes. Como resultado, la cadena de herramientas se fijó a una versión nightly exacta[7] para evitar que la implementación se rompiera antes de la publicación de la tesis.

Otro inconveniente digno de mención es que, en algunos casos, el código genérico podría adoptar la forma de una función cuyo comportamiento puede ser modelado por la misma red de Petri en todos los casos. En estas circunstancias, el MIR podría "condensarse" aún más antes de traducirlo a una red de Petri. Del mismo modo, algunas partes del MIR pueden ser superfluas para el análisis de detección de bloqueo y su traducción puede agrandar la salida, lo que ralentiza el análisis de alcanzabilidad realizado por el verificador de modelos. Esto puede contrarrestarse con optimizaciones cuidadosas que se propondrán en las Sec. 6.1 y 6.2.

3.3.3. Síntesis

En conclusión, a pesar de los inconvenientes mencionados anteriormente, interceptar la traducción en el nivel MIR ofrece ventajas significativas, entre las que se incluyen la maximización de la utilización del código del compilador existente, la reducción del esfuerzo de implementación y un mapeo más natural a las redes de Petri. Estas ventajas superan a los contras y hacen de la MIR un punto de partida convincente para la traducción en el contexto de la construcción de una herramienta de análisis estático para detectar deadlocks y señales perdidas en el código Rust.

Tanto [Meyer, 2020] como [Zhang and Liua, 2022] basan también sus traducciones en la MIR y, hasta donde sabe este autor, no existe ninguna herramienta análoga que realice una traducción a redes de Petri partiendo de una representación intermedia de nivel superior.

3.4. Mid-level Intermediate Representation (MIR)

An overview of the Mid-level Intermediate Representation (MIR) is provided in this section. MIR was introduced in RFC 1211[8] in August 2015. We will explore its different parts, how different code fragments are mapped to them, and the underlying graph structure.

```
1 // WARNING: This output format is intended for human consumers only
2 // and is subject to change without notice. Knock yourself out.
```

[7] https://github.com/Skasselbard/Granite/blob/master/rust-toolchain
[8] https://rust-lang.github.io/rfcs/1211-mir.html

```rust
fn main() {
    match std::env::args().len() {
        1 => 2,
        3 => 6,
        _ => 0,
    };
}
```

Listing 3.1: Simple Rust program to explain the MIR components.

```
fn main() -> () {
    let mut _0: ();                      // return place in scope 0 at src/main.rs:1:11: 1:11
    let mut _1: usize;                   // in scope 0 at src/main.rs:2:11: 2:33
    let mut _2: &std::env::Args;         // in scope 0 at src/main.rs:2:11: 2:33
    let _3: std::env::Args;              // in scope 0 at src/main.rs:2:11: 2:27

    bb0: {
        _3 = args() -> bb1;              // scope 0 at src/main.rs:2:11: 2:27
                                         // mir::Constant
                                         // + span: src/main.rs:2:11: 2:25
                                         // + literal: Const { ty: fn() ->
                                         //   Args {args}, val: Value(<ZST>) }
    }

    bb1: {
        _2 = &_3;                        // scope 0 at src/main.rs:2:11: 2:33
        _1 = <Args as ExactSizeIterator>::len(move _2) -> [return: bb2, unwind: bb4];
                                         // scope 0 at src/main.rs:2:11: 2:33
                                         // mir::Constant
                                         // + span: src/main.rs:2:28: 2:31
                                         // + literal: Const { ty: for<'a> fn(&'a Args) ->
                                         //   usize {<Args as ExactSizeIterator>::len},
                                         //   val: Value(<ZST>) }
    }

    bb2: {
        drop(_3) -> bb3;                 // scope 0 at src/main.rs:6:6: 6:7
    }

    bb3: {
        return;                          // scope 0 at src/main.rs:7:2: 7:2
    }
```

```
35
36      bb4 (cleanup): {
37          drop(_3) -> [return: bb5, unwind terminate]; // scope 0 at src/main.rs:6:6: 6:7
38      }
39
40      bb5 (cleanup): {
41          resume;                     // scope 0 at src/main.rs:1:1: 7:2
42      }
43  }
```

Listing 3.2: MIR of Listing 3.1 compiled using rustc 1.71.0-nightly in debug mode.

Consider the example code listed in Listing 3.1, the corresponding MIR[9] is shown in Listing 3.2. Notice the explicit warning at the top of the generated output. It will be omitted in the subsequent listings for simplicity. Moreover, output depends on the following factors:

- The *rustc* version in use, alternatively the release channel (stable, beta, or nightly).
- The build type: *debug* or *release*. By default, the command `cargo build` generates a *debug* build, while `cargo build -release` produces a *release* build.

To illustrate this variability, Listing 3.3 shows the output when compiling the same program in *release* mode. The distinguishing feature found in *release* builds is the presence of the `StorageLive` and `StorageDead` statements. On the other hand, *debug* builds generate shorter and clearer MIR that is closer to what the user wrote. For this reason, unless otherwise stated, the listings in this work contain MIR generated in *debug* builds.

```
1   // WARNING: This output format is intended for human consumers only
2   // and is subject to change without notice. Knock yourself out.
3   fn main() -> () {
4       let mut _0: ();                      // return place in scope 0 at src/main.rs:1:11: 1:11
5       let mut _1: usize;                   // in scope 0 at src/main.rs:2:11: 2:33
6       let mut _2: &std::env::Args;         // in scope 0 at src/main.rs:2:11: 2:33
7       let _3: std::env::Args;              // in scope 0 at src/main.rs:2:11: 2:27
8
9       bb0: {
10          StorageLive(_1);                 // scope 0 at src/main.rs:2:11: 2:33
11          StorageLive(_2);                 // scope 0 at src/main.rs:2:11: 2:33
12          StorageLive(_3);                 // scope 0 at src/main.rs:2:11: 2:27
13          _3 = args() -> bb1;              // scope 0 at src/main.rs:2:11: 2:27
14                                           // mir::Constant
15                                           // + span: src/main.rs:2:11: 2:25
16                                           // + literal: Const { ty: fn() ->
```

[9]The comments in the MIR have been slightly modified to improve the output

```
17                                  //   Args {args},
18                                  //   val: Value(<ZST>) }
19      }
20
21      bb1: {
22          _2 = &_3;               // scope 0 at src/main.rs:2:11: 2:33
23          _1 = <Args as ExactSizeIterator>::len(move _2) -> [return: bb2, unwind: bb4];
24                                  // scope 0 at src/main.rs:2:11: 2:33
25                                  // mir::Constant
26                                  // + span: src/main.rs:2:28: 2:31
27                                  // + literal: Const { ty: for<'a> fn(&'a Args) ->
28                                  //   usize {<Args as ExactSizeIterator>::len},
29                                  //   val: Value(<ZST>) }
30      }
31
32      bb2: {
33          StorageDead(_2);        // scope 0 at src/main.rs:2:32: 2:33
34          drop(_3) -> bb3;        // scope 0 at src/main.rs:6:6: 6:7
35      }
36
37      bb3: {
38          StorageDead(_3);        // scope 0 at src/main.rs:6:6: 6:7
39          StorageDead(_1);        // scope 0 at src/main.rs:6:6: 6:7
40          return;                 // scope 0 at src/main.rs:7:2: 7:2
41      }
42
43      bb4 (cleanup): {
44          drop(_3) -> [return: bb5, unwind terminate]; // scope 0 at src/main.rs:6:6: 6:7
45      }
46
47      bb5 (cleanup): {
48          resume;                 // scope 0 at src/main.rs:1:1: 7:2
49      }
50  }
```

Listing 3.3: MIR of Listing 3.1 compiled using rustc 1.71.0-nightly in release mode.

The specific formatting when converting MIR to a string has changed only slightly over time. See [Meyer, 2020, Section 3.3] for an example of older output from mid-2019.

As stated in Sec. 3.3, the MIR is derived from a previously existing control flow graph (CFG) in the Rust compiler. Fundamentally, a CFG is a graph representation of a program that exposes the underlying control flow.

3.4.1. MIR components

The MIR is formed by functions. Each function is represented as a series of basic blocks (BB) connected by directed edges. Each BB contains zero or more *statements* (usually abbreviated as "STMT") and lastly one *terminator statement*, for short *terminator*. The terminator is the only statement in which the program can issue an instruction that directs the control flow to another basic block inside the same function or to call another function. Branching as in Rust's `match` or `if` statements can occur only in terminators. Terminators play the role of mapping the high-level constructs for conditional execution and looping to the low-level representation in machine code as simple conditional or unconditional `branch` instructions.

In Fig. 3.1, the graph representation for the MIR shown in Listing 3.2 is presented as an example. The statements are colored in light blue and the terminators in light red. To make the kind of terminator statement clearer, extra annotations as in `CALL:` or `DROP:` were added.

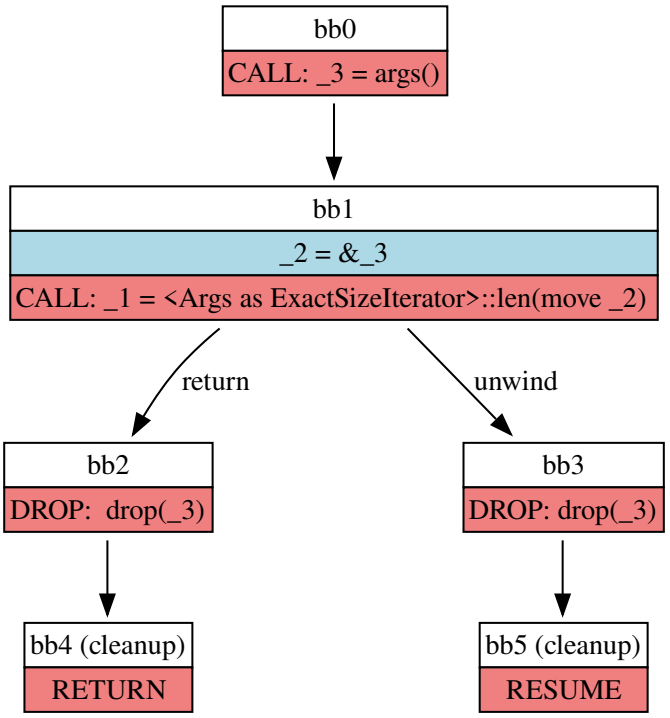

Figura 3.1: The control flow graph representation of the MIR shown in Listing 3.2.

It should be noted that the function call to `std::env::args().len()` in Line 2 in Listing 3.1 may return successfully or fail. A failure triggers an unwinding of the stack, ending the program and reporting an error. This is represented by the branching at the end of BB1 where the code execution may take the left path or the right path down the graph. The left branch (BB4 and BB5) corresponds to the correct execution of the program, while the right branch relates to the abnormal termination of the program.

There are different kinds of terminators and these are specific to the Rust semantics. We will introduce some of them to clarify the meaning of the example presented.

- As expected, a terminator of type `CALL:` calls a function, which returns a value, and continues execution to the next BB.
- A terminator of type `DROP:` frees up the memory of the variable passed in. It executes the destructors[10] and performs all the necessary cleanup tasks. From that point on, the variable cannot be used anymore in the program.
- `RETURN:` returns from the function. The return value is always stored in the local variable `_0`, as we will see shortly.
- `RESUME:` indicates that the process should continue unwinding. Analogously to a return, this marks the end of this invocation of the function. It is only permitted in cleanup blocks.

The complete list of terminator kinds can be found in the nightly documentation[11]. Other kinds of terminators will be discussed in detail in Sec. 4.4.3.

Regarding the variables, the data in MIR can be divided into two categories: *locals* and *places*. It is critical to observe that these "places" are *not* related to the places in Petri nets. Places are used to represent all types of memory locations (including aliases), while locals are limited to stack-based memory locations, i.e., local variables of a function. In other words, places are more general and locals are a special case of a place, therefore places are not always equivalent to locals. Conveniently, all the places are also locals in Fig. 3.1.

Locals are identified by an increasing non-negative index and are emitted by the compiler as a string of the form "`_<index>`". In particular, the return value of the function is always stored in the first local `_0`. This matches closely the low-level representation on the stack.

3.4.2. Step-by-step example

In this subsection, we will give a short explanation of what happens in each basic block of Fig. 3.1 to cover all the necessary information for the next sections. Moreover, this illustrates how the MIR output represents higher-level constructs often encountered when programming in Rust.

BB0

- The `main()` function starts at BB0.

[10] https://doc.rust-lang.org/stable/reference/destructors.html
[11] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/mir/enum.TerminatorKind.html

- A function is called (`std::env::args()`) to obtain an iterator over the arguments provided to the program.
- The return value of the function, the iterator, is assigned to the local `_3`.
- Execution continues in BB1.

BB1

- A reference to the iterator stored in `_3` is generated and stored in the local `_2` (similar to the "&" operator in C). This is necessary for calling methods because methods receive a reference to a struct of the same type (`&self`) as their first argument.
- The reference stored in `_2` is passed to the method `std::env::Args::len()` by moving and the function is called.
- The return value of the function, the number of arguments passed to the function, is assigned to the local `_1`.
- Execution continues in BB2 if successful, in BB4 in case of panic.

BB2

- The variable `_3`, whose value is the iterator over the arguments, is *dropped* since it is no longer needed.
- Execution continues in BB3.

BB3

- The function returns. The return value (local `_0`) is of type "unit"[12], which is similar to a `void` function in C, i.e., it does not return anything. This is how `main()` was defined in Listing 3.1.

BB4

- The variable `_3`, whose value is the iterator over the arguments, is *dropped* since it is no longer needed.
- If the drop is successful, execution continues in BB5, otherwise terminate the program immediately.

[12] https://doc.rust-lang.org/std/primitive.unit.html

BB5

- Continue unwinding the stack. This is the standard protocol defined for handling catastrophic error cases that cannot be handled by the program. Implementation details can be found in the documentation[13]

3.5. Function inlining in the translation to Petri nets

In this section, a thorough analysis and motivation for the third design decision listed at the beginning of the chapter, namely inlining function calls, is presented.

Modeling functions in PN is a crucial aspect of the translation because it is the basic unit of the MIR. By representing the functions in the MIR as PN and connecting them accordingly, the control flow and data shared between the threads in the program can be captured in a formal framework. Afterward, the Petri net is analyzed by a model checker in order to identify potential deadlocks or lost signals. This approach is especially useful when working with large and complex systems that may have many interrelated threads and functions, where the deadlock situation may not be evident even to an experienced code reviewer.

When translating MIR functions to PN, one key question that arises is whether to reuse the same representation for every call to a specific function or to "inline" the corresponding representation every time the function is called. Expressed differently, each function maps to a subnet in the final PN obtained after the translation, i.e., a connected subgraph formed by the places and transitions that model the behavior of the specific function. This smaller part of the net can either be present only once in the PN and all calls to this function connect to it, or be repeated for every instance of a call to the function in the Rust code.

Reusing the same model for every function seems at first glance more efficient, as the PN obtained is smaller. However, this approach can also lead to invalid states that were not present in the original Rust program. These can be the source of false positives during deadlock detection, as these extraneous states may violate the safety guarantees offered by the compiler.

On the other hand, inlining the model every time a function is called results in a larger PN, which requires more memory and CPU time to be analyzed, but it can also improve the accuracy of the analysis by ensuring that each function call is represented by a separate Petri net structure that captures its specific data dependencies in the context in which the function call occurs in the code.

[13] https://rustc-dev-guide.rust-lang.org/panic-implementation.html.

3.5.1. The basic case

The impact of these subtle details can only be fully comprehended with an appropriate example. Therefore, consider first the most simple abstraction of a function call in the language of Petri nets, formed by a single transition and two places representing the start and end of the function. This is seen in Fig. 3.2. The function call is treated as a black box, all details are abstracted away in the transition. We care only about where the function starts and where it ends.

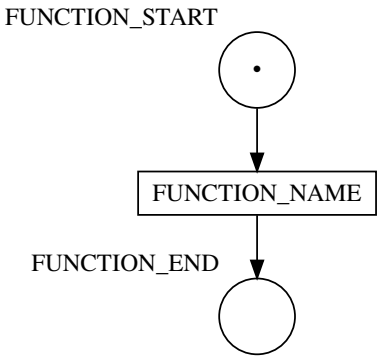

Figura 3.2: The simplest Petri net model for a function call.

Observe now such a function in the context of a Rust program. Listing 3.4 provides a simple example in which one function is called five times consecutively in a `for` loop. A possible PN that models the program is found in Fig. 3.3. It should be emphasized that this net and the subsequent ones in this section do *not* result from a translation of the MIR. They are simplifications to showcase the difficulties of dealing with functions called in various places in the code.

```
1 fn simple_function() {}
2
3 pub fn main() {
4     for n in 0..5 {
5         simple_function();
6     }
7 }
```

Listing 3.4: A simple Rust program with a repeated function call.

3.5.2. A characterization of the problem

The troublesome scenario has not emerged so far. It manifests only when a function is called in at least two different places in the code or, in simpler terms, the expression `simple_function()`

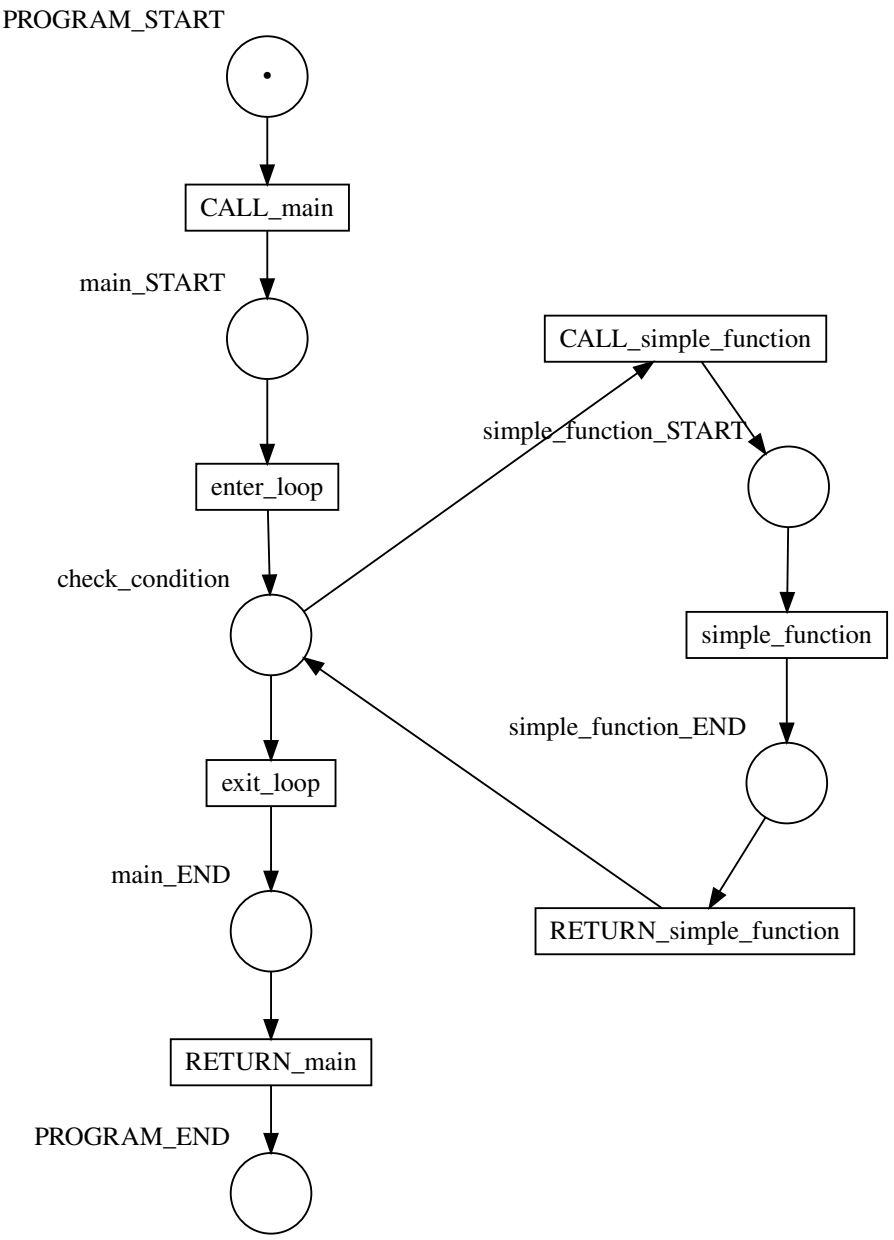

Figura 3.3: A possible Petri net for the code in Listing 3.4 applying the model of Fig. 3.2.

appears twice or more. Listing 3.5 satisfies this condition and is designed to exhibit the extraneous behavior described at the beginning of the section.

As stated before, the first approach to modeling the program consists in reusing the function model for both calls. This is shown in Fig. 3.4.

It is evident to the reader that the program in Listing 3.5 never calls the `panic!` macro and always terminates successfully, given that the variable `second_call` is never `true` before line 9.

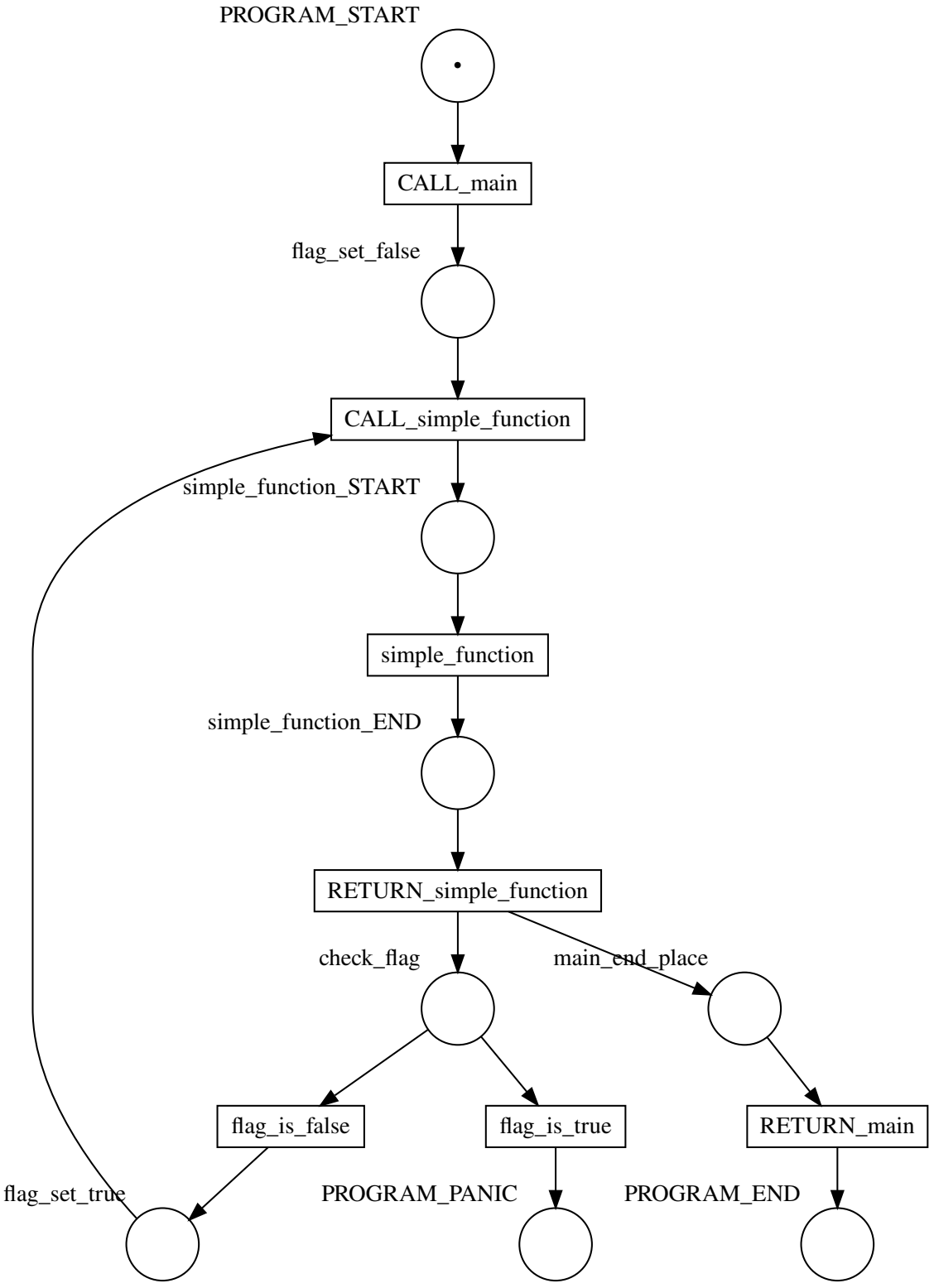

Figura 3.4: A first (incorrect) Petri net for the code in Listing 3.5.

```
1  fn simple_function() {}
2
3  pub fn main() {
4      let mut second_call = false;
5      simple_function();
6      if second_call {
7          panic!()
8      }
9      second_call = true;
10     simple_function();
11 }
```

Listing 3.5: A simple Rust program that calls a function in two different places.

Yet, the PN depicted in Fig. 3.4. is conspicuously flawed, making it unsuitable as a model for the program. The reason is that after firing the transition labeled `RETURN_simple_function` a token is placed in `check_flag` but *also* in `main_end_place`. The token in `main_end_place` will eventually appear in `PROGRAM_END`, which indicates a normal termination of the program. This is technically correct since we know that the program terminates successfully.

Nonetheless, there are concerning issues regarding the second token. The token in `check_flag` could be consumed either by the transition `flag_is_false` or `flag_is_true`. If it is consumed by the latter, a token will be placed in `PROGRAM_PANIC`, signaling an erroneous termination of the program. This is absurd because it means that the program could panic but also *always* ends normally, as seen in the previous paragraph.

The situation becomes worse if we follow the path of firing `flag_is_false`. In that case, the token triggers another function call, which is in principle correct, but nothing prevents it from doing this over and over again. The conclusion is that an infinite amount of tokens could accumulate in `main_end_place` or `PROGRAM_END` in the circumstance that, by pure chance, the transition `flag_is_true` does not fire.

It has become clear that we must discard this model and look for a better solution. One possibility is to split the transition labeled `RETURN_simple_function` in two separate transitions depending on the function call order as illustrated in Fig. 3.5.

This second attempt unfortunately comes with its own set of extraneous states. First, the program may now exit after calling the function only once. Nothing prevents the transition `RETURN_simple_function_2` from firing first. This is equivalent to saying that the execution flow jumps from line 5 to line 11 in Listing 3.5, which is obviously not a property present in the original Rust code.

On the other hand, the problem of the infinite loop persists. The PN may continue firing indefinitely as long as `flag_is_true` and `RETURN_simple_function_2` do not fire. There is no

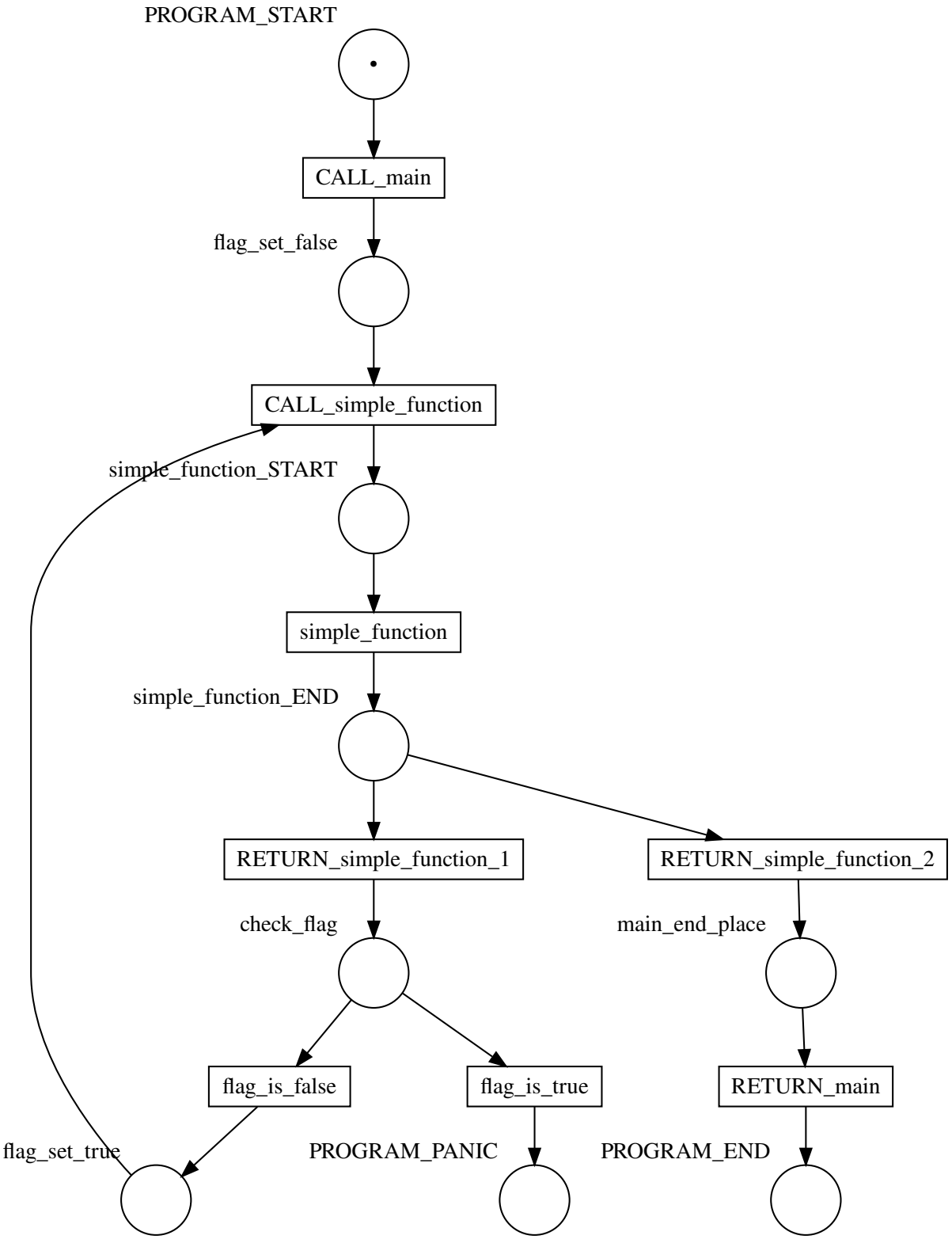

Figura 3.5: A second (also incorrect) Petri net for the code in Listing 3.5.

guarantee that the transitions fire in a specific order. As seen in Sec. 1.1.3, the transition firing is non-deterministic.

3.5.3. A feasible solution

Having observed the difficulties of modeling function calls, we turn our attention to the other approach to modeling function calls: Inlining the PN representation. Some of the lessons learned from the preceding subsection are:

- Creating a loop in the net where there is no loop in the original program opens the door to infinite sequences of transition firings. This could in turn break the *safety* property of the PN.
- As the token symbolizes the program counter, there must be only one token in the PN at any given time.
- The program state may change between function calls. Accordingly, separate places should model these states. Put differently, the state when calling a function the first time may not be the same as when calling the function a second time.

Fig. 3.6 introduces the inlining approach implemented in the tool. The PN therein is correct. It matches the structure of the Rust code more closely. It does not contain any loops nor it creates additional tokens when firing transitions, i.e., none of the transitions has two outputs. It is worth mentioning that the resulting PN is a state machine (Definition 8) as expected for a single-threaded program. This was not the case for Fig. 3.4 and 3.5.

A significant advantage of the inlining approach is that every function call is unequivocally identified. This proves helpful when interpreting the output of the model checker or error messages during the translation of a given program. The use of an incremental non-negative id is arbitrary but convenient. Moreover, the accuracy of deadlock detection is increased because certain classes of extraneous states such as those in the PN shown in the previous section are not present. Minimizing the number of false positives plays an important role when considering which approach to implement for a tool that aims to be user-friendly and easy to set up.

One disadvantage mentioned earlier is that the size of the resulting net is larger. The exact penalty in the number of additional places and transitions depends on the frequency with which functions are reused on average in the codebase. It is reasonable to assume that functions are called from several places. However, certain optimizations can be applied, which can reduce the size of the net considerably, thus compensating for the effect of using inlining. These optimizations are discussed in detail in Sec. 6.1 and 6.2.

Lastly, an attentive reader may notice that the analysis of the PN in Fig. 3.6 leads to the conclusion that the program may call `panic!` and terminate abruptly, which does not match the execution of the Rust program. This is correct but it is a limitation of low-level Petri nets that cannot be solved in the framework of the model and goes beyond the scope of this work. Sec. 6.6 explores the consequences of this restriction and proposes potential remedies.

Armed with new insights and knowledge about the design choices, we are now able to fully describe the implementation.

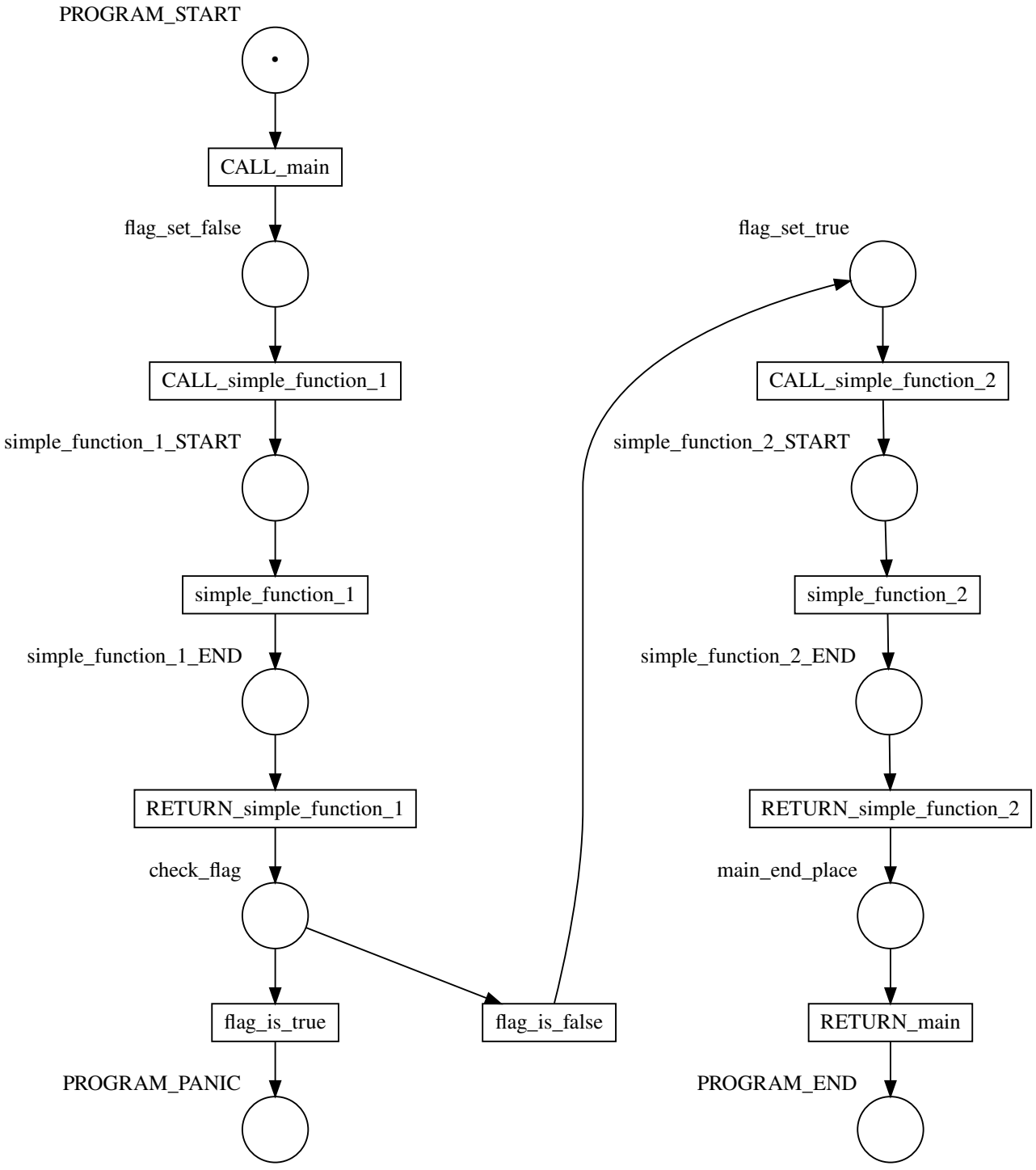

Figura 3.6: A correct Petri net for the code in Listing 3.5 using inlining.

Capítulo 4

Implementación de la traducción

This chapter is dedicated to exploring the implementation details of the deadlock detection tool. Its purpose is to provide a high-level view of the code and the data structures. The most important implementation decisions made throughout the development process are examined as well.

In the subsequent sections, we will describe the central components of the deadlock detection tool, including the internal representation of the call stack, the function memory model, and the translation of every constituent of a MIR function.

Later on, a significant portion of the discussion is devoted to explaining the support of multithreading and the modeling of synchronization primitives as Petri nets. Its implementation required careful design considerations to ensure correctness and efficiency.

The tool currently supports the following structures from the Rust standard library to synchronize access to shared resources and provide communication among threads:

- mutexes (`std::sync::Mutex`[1]),
- condition variables (`std::sync::Condvar`[2]),
- atomic reference counters (`std::sync::Arc`[3]).

While the main details are covered, this chapter is not intended to serve as a substitute for the code documentation. The code documentation in the form of comments, unit tests, and integration tests provides comprehensive information on the low-level specifics and usage of the tool. As stated before, the repository is publicly available on GitHub[4][5].

[1] https://doc.rust-lang.org/std/sync/struct.Mutex.html
[2] https://doc.rust-lang.org/std/sync/struct.Condvar.html
[3] https://doc.rust-lang.org/std/sync/struct.Arc.html
[4] https://github.com/hlisdero/cargo-check-deadlock
[5] https://github.com/hlisdero/netcrab

4.1. Initial considerations

4.1.1. Basic places of a Rust program

The basic Petri net model for a Rust program generated by the tool can be seen in Fig. 4.1. The place labeled `PROGRAM_START` contains a token and represents the initial state of the Rust program. This token will "move" from statement to statement and can thus be interpreted as the program counter of the CPU.

Correspondingly, the place labeled `PROGRAM_END` models the end state of the program after normal program termination, i.e., returning from the `main` function, regardless of the specific exit code. In other words, a main function that returns an error code because of invalid parameters or an internal program error is still considered a "normal" program termination. In other instances, however, the program may never reach this state if `main` never returns. These are known in Rust as "diverging functions"[6] and are supported by the tool.

Lastly, the place labeled `PROGRAM_PANIC` models the *abnormal* program termination, which happens when the program calls the `panic!` macro.

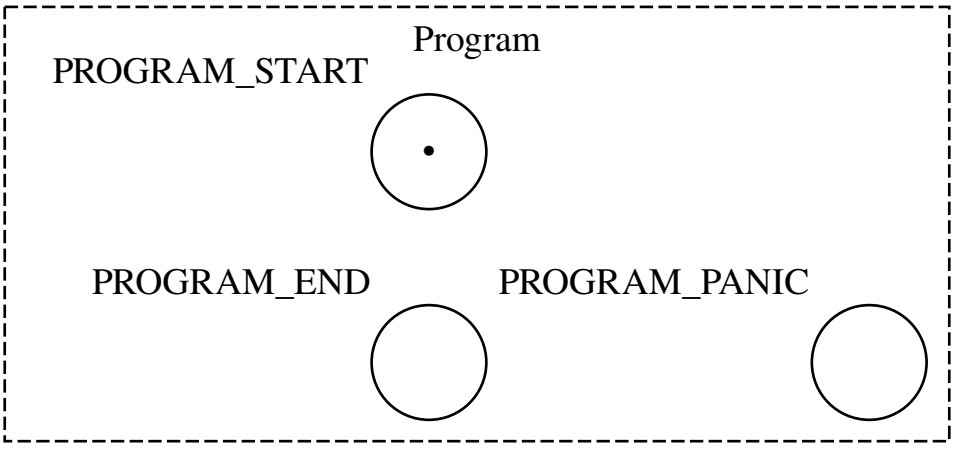

Figura 4.1: Basic places in every Rust program.

Two reasons for considering a separate panic end-state place can be argued. First, it is helpful for formal verification to distinguish the panic case from the normal termination case. A program may panic when detecting a possible violation of its memory safety guarantees. This is in most circumstances a wiser choice than simply ignoring the error and continuing. Therefore, such programs should not be flagged in principle as erroneous or defective but it is advisable to record the end state for troubleshooting and debugging purposes. Second, even if the user's code does not resort to `panic!` as an error-handling mechanism, numerous functions in the Rust standard

[6] https://doc.rust-lang.org/rust-by-example/fn/diverging.html

library may panic under extraordinary circumstances, e.g., due to out-of-memory (OOM) or hardware errors, or when the OS fails to allocate a new thread, mutex, etc. Consequently, it is essential to capture this eventual failure in the PN model.

There is one last subtle point that needs to be addressed. The program's start place is not as trivial as it seems. Although the `main` function is typically perceived as the first function to be executed, this is in reality not the case. Instead, Rust programs have a runtime that executes before the `main` function is called, in which language-specific features and static memory are initialized. One usually hears of interpreted languages such as Java or Python having a runtime but low-level languages such as Rust or C have a small runtime as well. It is simply thinner and less sophisticated. For interested readers, a guided tour of the journey before `main` was presented recently at a Rust conference [Levick, 2022].

Considering this, we are faced with the question of whether to include this runtime in the PN translation. On one hand, the runtime code is indeed part of the binary executed by the CPU. Nonetheless, it is platform-dependent code (the runtime is slightly different for every OS) and independent of the program's semantics, i.e., of the specific meaning of the program the user wrote. Since the user does not have any influence on this part of the binary, synchronization problems cannot be attributed to him/her. As such, this code does not add value to the translation and can be safely abstracted away, reducing in the process the size of the PN. In conclusion, the decision is to skip the runtime code; the translation starts at the `main` function.

4.1.2. Argument passing and entering the query

The tool is designed around a simple command-line interface (CLI). After parsing the command-line arguments using the well-known library `clap` library[7], the program enters a query to the *rustc* compiler to start the translation process. The majority of the work from that point on is coordinated by the struct of type `Translator`[8].

The query system was described briefly in Sec. 3.2. Two examples of the use of this mechanism are provided in the documentation[9][10]. They have proved extremely useful as a starting point, since they provide an excellent short working example of how to interact with *rustc*. In simpler terms, they are the "Hello, World!" of working side-by-side with the Rust compiler.

[7] https://docs.rs/clap/latest/clap/
[8] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator.rs
[9] https://rustc-dev-guide.rust-lang.org/rustc-driver-interacting-with-the-ast.html
[10] https://rustc-dev-guide.rust-lang.org/rustc-driver-getting-diagnostics.html

4.1.3. Compilation requirements

As briefly mentioned in Sec. 3.2.2, the tool must be compiled with the nightly version of *rustc* to access its internal crates and modules. The decisive section in the file *lib.rs*[11] is depicted in Listing 4.1.

```
13 // This feature gate is necessary to access the internal crates of the compiler.
14 // It has existed for a long time and since the compiler internals will never be
   ↪  stabilized,
15 // the situation will probably stay like this.
16 // <https://doc.rust-lang.org/unstable-book/language-features/rustc-private.html>
17 #![feature(rustc_private)]
18
19 // Compiler crates need to be imported in this way because they are not published on
   ↪  crates.io.
20 // These crates are only available when using the nightly toolchain.
21 // It suffices to declare them once to use their types and methods in the whole crate.
22 extern crate rustc_ast_pretty;
23 extern crate rustc_const_eval;
24 extern crate rustc_driver;
25 extern crate rustc_error_codes;
26 extern crate rustc_errors;
27 extern crate rustc_hash;
28 extern crate rustc_hir;
29 extern crate rustc_interface;
30 extern crate rustc_middle;
31 extern crate rustc_session;
32 extern crate rustc_span;
```

Listing 4.1: Excerpt of the file *lib.rs* showcasing how to use the *rustc* internals.

The `rustc_private` is a feature flag that controls access to the compiler's private crates. These crates are not installed by default when installing the Rust toolchain using *rustup*[12]. Hence, it is necessary to install the additional components *rustc-dev*, *rust-src*, and *llvm-tools-preview*. The purpose of each component is detailed in [Rust Project, 2023d]. Straightforward instructions to set up a development environment are also found in the `README`[13] of the repository.

To the best knowledge of this author, an alternative method of accessing the internals of the

[11] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/lib.rs
[12] https://rustup.rs/
[13] https://github.com/hlisdero/cargo-check-deadlock/blob/main/README.md

Rust compiler does not exist. Tools such as Clippy[14] or Kani [15], or kernels like Redox[16] and RustyHermit[17] use this mechanism as well.

4.2. Function calls

4.2.1. The call stack

A Rust program is composed, as in other programming languages, of functions. The program begins (except for the caveats seen in Sec. 4.1.1) with a call to the `main` function, which then may call other functions. It should be emphasized that function calls may be placed at any point within the code. A function can be called from another function or even from within itself, resulting in recursive calls.

Function calls are stored in memory in a data structure called the *call stack*. When a function is called in Rust, it gets pushed onto the call stack, creating a new stack frame. A stack frame contains important information such as the function's local variables, arguments, and the return address indicating where the program should resume once the function finishes its execution.

The call stack operates based on the principle of last in, first out (LIFO). As functions are called, each new stack frame is placed on top of the previous one. This allows the program to execute the most recently called function first. Once a function completes its execution, it is popped off the stack, and the program continues from the point where it left off in the previous function.

Hence, the call stack plays an essential role in managing function calls and returns, since it keeps track of the flow of function calls and maintains the necessary information for the program to return to the correct execution point after a function completes its task.

For the same reasons, mirroring the call stack in the translator is the most suitable approach for tracking function calls to be translated because it aligns with the logical flow of program execution. As functions are translated, they are pushed and popped from the call stack of the `Translator`, reflecting the order in which they are called at runtime. This enables us to handle nested function invocations and follow the control flow from one function to the other during the translation process.

[14] https://github.com/rust-lang/rust-clippy/blob/master/rust-toolchain
[15] https://github.com/model-checking/kani/blob/main/rust-toolchain.toml
[16] https://gitlab.redox-os.org/redox-os/redox/-/blob/master/rust-toolchain.toml
[17] https://github.com/hermitcore/rusty-hermit/blob/master/rust-toolchain.toml

4.2.2. MIR functions

In the implementation, the `Translator` has a stack that supports the usual operations `push`, `pop`, and `peek`. This stack stores structures of type `MirFunction`[18]. Later, we will see that not all functions are translated as MIR functions since not all functions have a representation in MIR and, in other cases, it is convenient to handle them differently. Nevertheless, MIR functions are the "common case" in the translation process, the default case for the majority of user-defined functions.

The available interface provided by *rustc* allows for querying the MIR body of only one function at a time, which can be done using the `optimized_mir`[19] method. This implies that it is not possible to get the MIR of the whole program initially and the translator must obtain the MIR from each function as it reaches them in the code. But how to identify each function? It is known from experience that functions in distinct modules may have the same name, making the name unsuitable as an identifier. Luckily, this problem is already solved in the compiler. The functions are uniquely identified by the compiler type `rustc_hir::def_id::DefId`[20]. This ID is valid for the crate currently being compiled and it is already present in the HIR. The high-level algorithm can be described as follows.

When the translation starts:

1. Query the id of the entry point of the program (the `main` function).
2. Create a `MirFunction` with the necessary information.
3. Push it to the stack.
4. If necessary, modify the MIR function contents using `peek`.
5. Translate the top of the call stack.
6. When `main` finishes, remove it (`pop`) from the call stack.

When a terminator of type "call" (see Sec. 3.4.1) is encountered:

1. Query the id of the called function.
2. Create a `MirFunction` with the necessary information.
3. Push it to the stack.
4. If necessary, modify the MIR function contents using `peek`.
5. Translate the top of the call stack.

[18] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/mir_function.rs

[19] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/ty/context/struct.TyCtxt.html#method.optimized_mir

[20] https://doc.rust-lang.org/stable/nightly-rustc/rustc_hir/def_id/struct.DefId.html

6. When the function finishes, remove it (`pop`) from the call stack.

As seen, the approach is consistent for every MIR function and thus is easier to implement.

The use of a call stack in the translation process enables context switching between MIR functions and facilitates the ability to return to the specific basic block from which a function was called. This allows for the translation of the program to be performed function by function, in a linear fashion, ensuring that the structure and order of the original program are maintained.

However, employing the call stack approach does come with certain limitations. First, if the same function is called multiple times within the program, it will be translated multiple times as well. This is related to the inlining strategy elaborated in Sec. 3.5. Although this can potentially be mitigated through the use of some sort of cache, it is out of the scope of this thesis. This optimization will be discussed in Sec. 6.3.

The more severe implication of using the call stack approach is the inability to handle recursive functions. When encountering a recursive function within the translation process, the process becomes trapped in an endless loop where the stack grows indefinitely as new stack frames are pushed to it, leading to a stack overflow and a subsequent crash of the translation process. This problem is addressed in Sec. 6.4 too. For now, it is necessary to accept the limitation that recursive functions cannot be translated using this framework.

4.2.3. Foreign functions and functions in the standard library

In Rust, the compiler includes by default the standard library in all compiled binaries, effectively linking it statically. To override this behavior, the crate-level attribute `#![no_std]` is used to indicate that the crate will link to the core-crate instead of the std-crate. See [Rust on Embedded Devices Working Group, 2023] for more details.

This means that the standard library's functionality becomes an integral part of the resulting executable. Function calls to the standard library appear in various contexts in Rust code, such as when accessing command line arguments, invoking iterators, utilizing traits like `std::clone::Clone`, `std::deref::Deref::deref`, or employing standard library types like `std::result::Result` or `std::option::Option`. Given the prevalence of these function calls throughout Rust programs, it becomes essential to handle them separately in the translation process. It is evident that these standard library functions, due to their purpose, cannot lead to a deadlock. Therefore, it is more practical to treat them as black boxes within the translation process, bypassing the need to translate their MIR. This approach is indispensable in order to avoid generating an excessively large and convoluted Petri net that would hinder readability and comprehension.

The focus of the translation effort lies primarily on the user code, specifically the functions that developers write to implement their desired functionalities. By directing attention to the user code and excluding the translation of standard library functions, the resulting Petri net

remains more manageable, facilitating the analysis and verification of potential deadlocks within the user's codebase. The calls to the standard library constitute, in other words, the "frontier" or "boundary" of the translation, the point at which we stop translating the MIR accurately and rely instead on a simplified model.

Petri net model for a function with cleanup block

The model presented in Fig. 3.2 is the first approximation. There is, however, an implementation detail that requires careful attention. Numerous functions in the standard library contain not only an end place ("target block", in the *rustc* parlance) but also a cleanup place ("cleanup block"). This second execution path is taken when the function explicitly panics or more generally fails to achieve its goal for whatever reason. In this case, the control flow continues to a different basic block, where variables are freed and eventually the program ends with a panic error code. Stated differently, the unwind of the stack begins as soon as a function encounters a non-recoverable failure.

Considering that the translator cannot tell if this abnormal situation could lead to a deadlock later in the translation process, it is imperative to translate this alternative execution path whenever possible. Only in counted exceptions, all related to the synchronization primitives and discussed in the respective sections, this cleanup block is ignored explicitly. The complete model for an abridged function call with a cleanup block can be seen in Fig. 4.2.

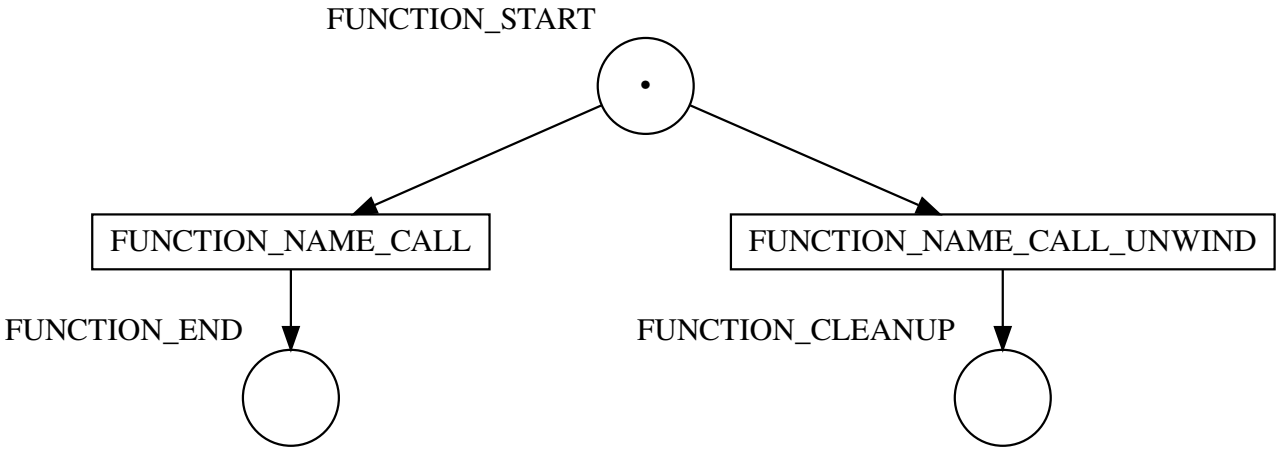

Figura 4.2: The Petri net model for a function with a cleanup block.

Functions translated with the abridged Petri net model

Having discussed the exclusion of standard library functions from the translation process, we now shift our focus towards the functions that indeed require translation using the model we presented earlier. Surprisingly, they include a considerable number of functions.

- Functions part of the standard library (the std-crate[21]), save for the `std::sync::Condvar::wait` function detailed in Sec. 4.8.3.
- Functions part of the core library (the core-crate[22]).
- Functions in the alloc-crate: the core allocation and collections library[23].
- Functions without a MIR representation This can be checked with the `is_mir_available` method[24].
- Functions which are a foreign item i.e., linked via `extern { ... }`. This can be checked with the `is_foreign_item` method[25].

In the future, calls to functions in dependencies, i.e., in other crates, should also be handled in this fashion. In conclusion, the default case for functions that are *not* user-defined is to treat them as a foreign function and use an abridged Petri net model to translate them.

4.2.4. Diverging functions

Diverging functions are a special case that is relatively easy to support. It is simply a function that never returns to the caller. Examples of this are a wrapper around an infinite `while` loop, a function that exits the process, or a function that starts an OS. It suffices to connect the start place of the function to a sink transition (Definition 7) as seen in Fig. 4.3.

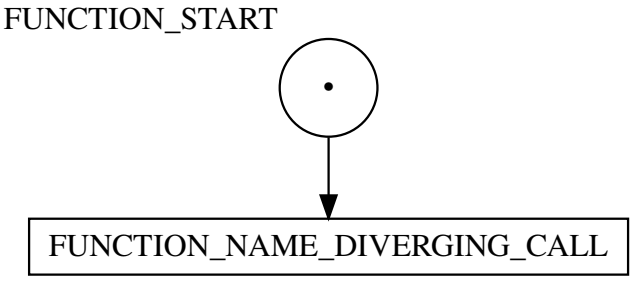

Figura 4.3: The Petri net model for a diverging function (a function that does not return).

Note that this special case does not constitute a deadlock and must *not* be treated as such. An infinite loop, i.e., a "busy wait", is in its inherent nature distinct from the infinite wait that characterizes a deadlock as seen in Sec. 1.4.1. In other words, detecting infinite loops is closer to

[21] https://doc.rust-lang.org/std/
[22] https://doc.rust-lang.org/core/
[23] https://doc.rust-lang.org/alloc/
[24] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/ty/context/struct.TyCtxt.html#method.is_mir_available
[25] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/ty/context/struct.TyCtxt.html#method.is_foreign_item

the problem of detecting livelocks, which are out of the scope of this thesis. Besides, the translator cannot tell ahead of time if the diverging call is benign like a call to `std::process::exit` or a call to some kind of function carefully designed to block the program.

In the current PN model, the token is consumed and the net is left in an end state without tokens in the places `PROGRAM_END` or `PROGRAM_PANIC` shown in Fig. 4.1. Consequently, the model checker is able to distinguish this end state from the other cases and conclude that a diverging function has been called.

4.2.5. Explicit calls to panic

The `panic!` macro can be seen as a special case of a divergent function where the transition representing the function call is connected to the place labeled `PROGRAM_PANIC` described in Sec. 4.1.1. The translator detects an explicit call to panic, which is one of the following functions:

- `core::panicking::assert_failed`
- `core::panicking::panic`
- `core::panicking::panic_fmt`
- `std::rt::begin_panic`
- `std::rt::begin_panic_fmt`

The documentation[26] elaborates on why panic is defined in the core-crate and the std-crate and how it is implemented.

See Listing 4.2 for a simple program that panics. The corresponding Petri net model is depicted in Fig. 4.4. This is one of the illustrative examples included in the repository.

```
1 fn main() {
2     panic!();
3 }
```

Listing 4.2: A simple Rust program that calls `panic!`.

4.3. MIR visitor

This section is dedicated to a pivotal component that serves as the backbone of the translator: the MIR Visitor trait[27]. This trait facilitates straightforward navigation of the MIR of the Rust

[26] https://rustc-dev-guide.rust-lang.org/panic-implementation.html
[27] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/mir/visit/trait.Visitor.html

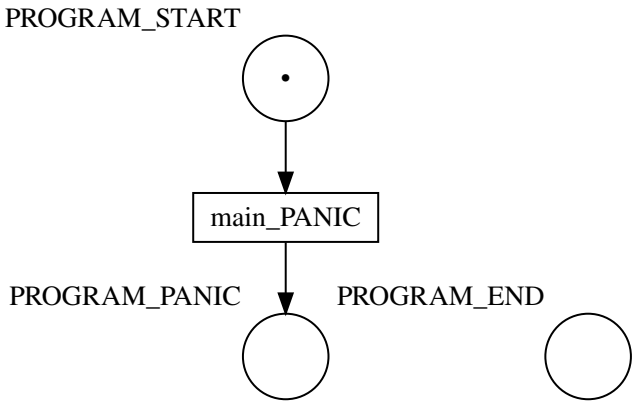

Figura 4.4: The Petri net model for Listing 4.2.

source code. In other words, it acts as the glue that seamlessly binds the various components of the translator together.

The MIR Visitor trait plays a fundamental role in the translation process by providing a structured approach to traverse and analyze the MIR. It offers a set of methods that can be implemented to perform specific actions at different points during the traversal. By employing this trait, the translator gains the ability to systematically explore the MIR and extract the necessary information for generating the corresponding Petri net.

The implemented methods within the MIR Visitor trait serve as entry points for handling different elements encountered during the traversal. These methods allow for customized processing of specific MIR constructs, e.g., basic blocks, statements, terminators, assignments, constants, etc. By defining appropriate behavior for each method, the translator can efficiently extract relevant data and make informed decisions based on the encountered MIR elements.

It was not required to implement all possible methods. If not defined, the methods in MIR Visitor simply call the corresponding `super` method and continue the traversal. For instance, `visit_statement` calls `super_statement` if no custom implementation is present. In the case of the translator, the implemented methods are:

- `visit_basic_block_data` for keeping track of the basic block currently being translated.
- `visit_assign` for keeping track of assignments of synchronization variables (mutexes, mutex guards, join handles, and condition variables).
- `visit_terminator` for processing the terminator statement of each basic block, that is, connecting the basic blocks.

To start visiting the MIR, the method `visit_body` must be used. Listing 4.3 shows the corresponding function in the translator.

In conclusion, the MIR Visitor trait simplifies remarkably the translation as it is not necessary to implement a traversal mechanism thanks to the provided compiler interfaces. This also makes

```rust
/// Main translation loop.
/// Translates the function from the top of the call stack.
/// Inside the MIR Visitor, when a call to another function happens, this method will be called again
/// to jump to the new function. Eventually a "leaf function" will be reached, the functions will exit and the
/// elements from the stack will be popped in order.
fn translate_top_call_stack(&mut self) {
    let function = self.call_stack.peek();
    // Obtain the MIR representation of the function.
    let body = self.tcx.optimized_mir(function.def_id);
    // Visit the MIR body of the function using the methods of `rustc_middle::mir::visit::Visitor`.
    //
    <https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/mir/visit/trait.Visitor.html>
    self.visit_body(body);
    // Finished processing this function.
    self.call_stack.pop();
}
```

Listing 4.3: The method in the `Translator` that starts the traversal of the MIR.

the translator more robust and resistant to changes in *rustc*. If the string representation of the MIR changes, the translator is left unaffected. As long as the internal interfaces for accessing the MIR stay the same, the translator can navigate the MIR semantically and not based on how it is printed to the user.

As a last remark, similar traits exist for other intermediate representations:

- AST: https://doc.rust-lang.org/stable/nightly-rustc/rustc_ast/visit/trait.Visitor.html
- HIR: https://doc.rust-lang.org/stable/nightly-rustc/rustc_hir/intravisit/trait.Visitor.html
- THIR: https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/thir/visit/trait.Visitor.html

Various components in the compiler implement these traits to navigate the intermediate representations. To put it roughly, they are analogous to iterators for collections.

4.4. MIR function

In the following section, we will delve into the translation process of a MIR function. This section aims to provide a comprehensive understanding of the translation techniques applied to specific MIR elements, namely basic blocks (BB), statements, and terminators. These components were introduced previously in Sec. 3.4.1.

The implementation in the repository is accordingly named `MirFunction`[28]. This type stores the start place and the end place of the function. These must be supplied to the MIR function because they also represent where the function call took place and where it should return to. The end place is in simpler terms the return place in the Petri net. See Fig. 3.2 for an illustration.

The start place of the function overlaps with the place that models the first basic block in the function. This matches more closely the MIR as the code only lives inside of basic blocks, so the function call begins at the first basic block (BB0).

The `MirFunction` also stores the ID that identifies it. This is necessary for performing function calls from this function. Moreover, the function requires a name that is different for every function call, so it receives a name with an index appended, making it unique across the whole Petri net.

We will now explain how each component is expressed in the language of Petri nets. Through a detailed exploration of the translation techniques employed for basic blocks, statements, and terminators, we will develop a formal model that accurately captures the behavior of a MIR function for deadlock detection.

4.4.1. Basic blocks

One aspect of the translation process involves transforming basic blocks into Petri nets, which serve as a fundamental building block for modeling the control flow within the MIR function. As seen in Fig. 3.1, a basic block in MIR acts as a container that houses a sequence of zero or more statements, as well as a mandatory terminator statement.

As nodes in a graph, the main property of basic blocks is their ability to direct the flow of control within a program. Each basic block may have one or more basic blocks pointing to it, indicating the potential paths from which the control flow can reach it. Similarly, a basic block can point to one or more other basic blocks, signifying the possible paths the control flow may take after executing the current basic block. It is worth mentioning that isolated basic blocks with no connections do not make sense since they would never be executed, i.e., they are dead code.

[28] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/mir_function.rs

This branching behavior allows for dynamic control flow within the program, as multiple basic blocks can continue the control flow to the same target basic block (for instance to a block that performs cleanup tasks). Conversely, a basic block can branch and determine the next basic block based on specific conditions or program logic, e.g., in an `if`, `while`, `match` or other control structures. This versatility in control flow provides the foundation for modeling complex program behavior.

The Petri net model used in the implementation relies on a single place to model each BB. We can abstract away the inner workings of the BB and work with a single place. The rationale behind it is that the connections to other BB depend solely on the terminators and statements are not modeled at all as we will see shortly. Additionally, the implementation[29] keeps track of the function name to which the BB belongs and the BB number to generate unique labels.

4.4.2. Statements

MIR statements are intentionally *not* incorporated into the Petri net model. Considering the reasons for this and the benefits may not be immediately apparent, we will provide a detailed explanation for this implementation decision.

The approach that was previously implemented did include the modeling of statements. It was based on the approach seen in [Meyer, 2020]. However, it was observed that this led to the creation of a long chain of places and transitions that did not significantly contribute to the detection of deadlocks or missed signals. Furthermore, it unnecessarily inflated the size of the Petri net representation, making it more difficult to debug and understand. Consequently, this approach was later revised and removed in a later commit[30].

In all the programs we had tested so far, the statements did not perform any action that may justify their addition to the Petri net. On the contrary, the nets that include the statements were larger and more difficult to read. In order to facilitate the use and adoption of the tool, it is crucial to optimize the Petri net for the purpose of the tool. The decision was therefore to eliminate all the code related to the modeling of the statements and fix the tests accordingly to match the new output.

The alternative was to disable the statements with a compile flag but that would complicate testing and since there is no use case for modeling the MIR statements anyway, this option was discarded.

For illustrative purposes, we can refer to Fig. 4.5, which showcases a comparison between the old model and the new model. The differences between these two representations are evident, high-

[29] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/mir_function/basic_block.rs

[30] https://github.com/hlisdero/cargo-check-deadlock/commit/b27403b6a5b2bb020a5d7ab2a9b1cacefb48be82

lighting the removal of statements from the model and the subsequent simplification achieved in the resulting Petri net.

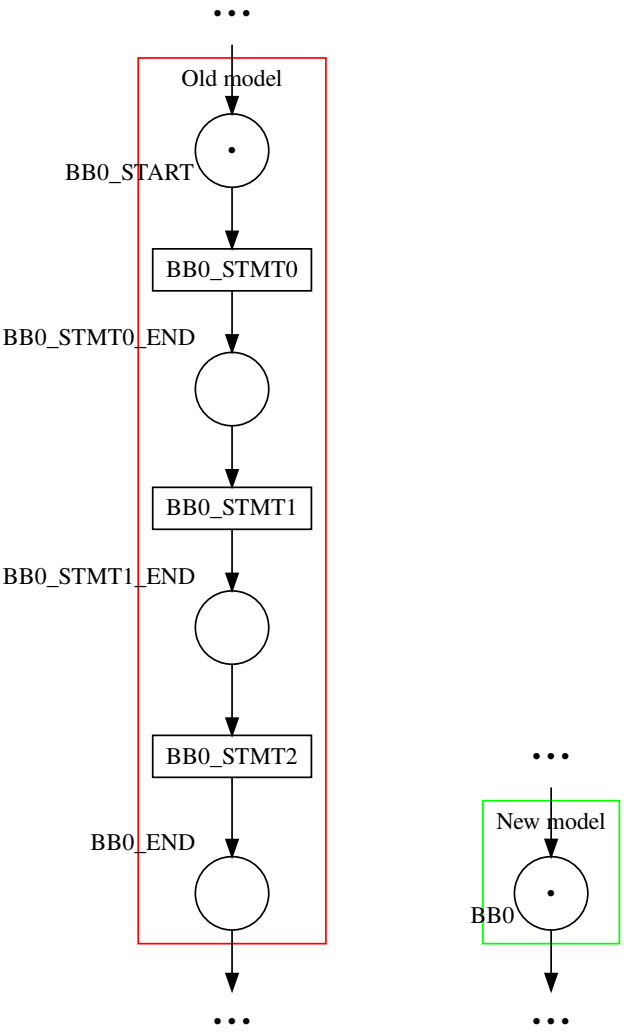

Figura 4.5: A side-by-side comparison of two possibilities to model the MIR statements.

4.4.3. Terminators

As seen in Sec. 3.4.1, terminator statements come in different shapes. The documentation for the enum `TerminatorKind`[31] lists as of this writing 14 different variants. The implementation is required to support most of them, since they appear sooner or later in the test programs included in the repository and their translation directly influences the connections between

[31] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/mir/enum.TerminatorKind.html

the basic blocks. The remaining terminators that are not implemented are not present when querying the `optimized_mir`, i.e., they are used only in previous compiler passes.

The implementation of the MIR Visitor[32] includes the `visit_terminator` method as seen before. This is where the edges connecting one BB to another are created. In the next paragraphs, the high-level details of each handler are discussed. Some implementation details are omitted as they do not affect the Petri net.

Goto

This is an elementary terminator kind. The end place of the currently active BB is connected to the start place of the target BB through a new transition with an appropriate label.

SwitchInt

This terminator kind comes with a collection of target basic blocks. For each target BB, we connect the end place of the currently active BB to the start place of the target BB through a new transition with an appropriate label. This creates a *conflict* as defined in Sec. 1.1.3.

The label must also contain some kind of unique identifier of the block from where the jump starts. This is a precondition to correctly translate multiple basic blocks with a `SwitchInt` that jumps to the same block.

Resume or Terminate

These are terminators that model respectively an unwinding of the stack and the immediate abort of the program. Both are treated in the same way: Connecting the end place of the currently active BB to the `PROGRAM_PANIC` place seen in Fig. 4.1.

Return

This is the terminator that causes the MIR function to return. This is where the end place of the function is used. The end place of the currently active BB is connected to it.

Unreachable

This is a border case that appears in some `match`, `while` loops, or other control structures. The documentation states: *Indicates a terminator that can never be reached.* To handle this case,

[32] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/mir_visitor.rs

the decision was to connect the end place of the currently active BB to the `PROGRAM_END` place seen in Fig. 4.1. See the comments in the repository for more details.

Drop

The `std::ops::Drop` trait is used to specify code that should be executed when the type goes out of scope [Klabnik and Nichols, 2023, Chap. 15.3]. It is equivalent to the concept of destructors found in other programming languages.

The drop terminator behaves like a function call with a cleanup transition. Therefore, we apply the model shown in Fig. 4.2 with modified transition labels.

An important check happens here too, namely checking if a mutex guard is being dropped. If that is the case, then the corresponding mutex should be unlocked as part of the transition firing. Precise details are explained in Sec. 4.7.3.

Call

This is the terminator kind for executing function calls. The presence of a cleanup block and the particular `UnwindAction`[33] as well as the function name and type is analyzed to handle it according to the strategy elaborated in Sec. 4.2.

Note that `UnwindAction` is a refactor of *rustc* that was introduced on April 7th, 2023. It is a good example of a regression that required significant changes to accommodate. The interested reader is referred to the corresponding commit[34].

Assert

This terminator kind is related to the `assert!()`[35] macro and the default overflow checks that *rustc* incorporates when performing arithmetic operations.

The implementation does not model the condition for the assert. It simply connects the end place of the currently active BB to the start place of the target BB through a new transition with an appropriate label.

In some cases, a cleanup block is present too. For this, a second transition is needed, analogously to the `Drop` case.

[33] https://doc.rust-lang.org/stable/nightly-rustc/rustc_middle/mir/syntax/enum.UnwindAction.html

[34] https://github.com/hlisdero/cargo-check-deadlock/commit/8cf95cd54b29c210801cae2941abcbbb85051b92

[35] https://doc.rust-lang.org/std/macro.assert.html

4.5. Function memory

We will now proceed to explore the memory characteristics of the MIR function in detail. It is important to acknowledge that the need to record values being assigned between memory locations in the MIR arises from the requirements of the deadlock and missed signals detection. In simpler teams, we are forced to model the memory only because the supported synchronization variables need to be tracked during the translation process.

The translator must track variables of the following types:

- Mutexes (`std::sync::Mutex`).
- Mutex guards (`std::sync::MutexGuard`).
- Join handles (`std::thread::JoinHandle`).
- Condition variables (`std::sync::Condvar`).
- Aggregates, i.e., wrappers such as `std::sync::Arc` or types that contain multiple values like tuples or a structured type (`struct`).

Before calling methods on these types of synchronization variables, immutable or mutable references to the original memory location are created. The translator must somehow know which specific synchronization variable is behind a given reference. Knowing the type of the memory location is *not* enough, the *value* must be readily available to the translator to operate on the Petri net model of the specific synchronization variable.

4.5.1. A guided example to introduce the challenges

To illustrate the situation described previously, consider the Rust program shown in Listing 4.4. It is again one of the example programs found in the repository. As it should be evident to the reader, this program deadlocks when executed. The reason is that the `std::sync::Mutex::lock` method is being called twice on the same mutex. To detect this deadlock, the translator must be able to at the very least identify that the invocation of `lock` takes place on the same mutex.

```
1 fn main() {
2   let data = std::sync::Mutex::new(0);
3   let _d1 = data.lock();
4   let _d2 = data.lock(); // cannot lock, since d1 is still active
5 }
```

Listing 4.4: A deadlock caused by calling `lock` twice on the same mutex.

Observe now an excerpt of the MIR of the same erroneous program in Listing 4.5. The comments have been removed for clarity. In BB0 the mutex is created through a call to `std::sync::Mutex::new`.

The new mutex is the return value of the function. It is assigned to the local variable `_1`. Then the execution continues in BB1. Focus on the first statement of BB1: An immutable reference to the local variable `_1` is stored in `_3`. Next, the reference is moved to the function `std::sync::Mutex::lock`. This reference is consumed by lock, that is to say, the local variable `_3` is not used anywhere else in the MIR because, from that point on, the ownership of the reference is transferred to the `std::sync::Mutex::lock` function.

```
fn main() -> () {
  let mut _0: ();
  let _1: std::sync::Mutex<i32>;
  let mut _3: &std::sync::Mutex<i32>;
  let mut _5: &std::sync::Mutex<i32>;
  scope 1 {
      debug data => _1;
      let _2: std::result::Result<std::sync::MutexGuard<'_, i32>,
      ↪ std::sync::PoisonError<std::sync::MutexGuard<'_, i32>>>;
      scope 2 {
          debug _d1 => _2;
          let _4: std::result::Result<std::sync::MutexGuard<'_, i32>,
          ↪ std::sync::PoisonError<std::sync::MutexGuard<'_, i32>>>;
          scope 3 {
              debug _d2 => _4;
          }
      }
  }

  bb0: {
      _1 = Mutex::<i32>::new(const 0_i32) -> bb1;
  }

  bb1: {
      _3 = &_1;
      _2 = Mutex::<i32>::lock(move _3) -> bb2;
  }

  bb2: {
      _5 = &_1;
      _4 = Mutex::<i32>::lock(move _5) -> [return: bb3, unwind: bb6];
  }
```

Listing 4.5: An except of the MIR of the program from Listing 4.4.

Immediately after the statement, the translator encounters the terminator of BB1. It contains

a call to `std::sync::Mutex::lock`. How would the translator know, when translating this call, that `_3` is indeed the mutex stored in `_1`? This is the problem that the modeling of the function's memory aims to solve.

The problem goes even further. The local variable `_2` contains a mutex guard after the call to lock, which should be recorded too. Notice how BB2 repeats the same operations as BB1 but uses different local variables, `_5` and `_4`. The translator should know that `_5` is an alias for `_1` as well. Furthermore, the mutex guards in `_2` and `_4` will eventually be dropped, which indirectly unlocks the mutex. There has to be a link from the mutex guard in `_2` and `_4` to the mutex in `_1`. More concisely, the translator should monitor which mutex is behind each mutex guard.

To make matters more complex, each MIR function has its own stack memory, with its separate local variables `_0`, `_1`, `_2`, `_3`, and so on. Thus, the mapping of memory locations to synchronization variables cannot be a single global structure. It is instead dependent on the context of the current function being translated. Lastly, a synchronization variable could migrate from one function to another and the translator must be able to re-map them correctly.

This suffices as a brief practical example of the challenges of memory modeling. We can now introduce the solution that has been implemented.

4.5.2. A mapping of `rustc_middle::mir::Place` to shared counted references

The implementation is suitably named `Memory`[36]. As anticipated in the previous section, there is one instance of `Memory` per `MirFunction`. The memory is tightly connected to the context of the MIR function.

Rather than moving values between different memory locations, as observed in the MIR, our solution relies on the simpler concept of "linking". This entails associating a specific `rustc_middle::mir::Place` with the corresponding value. This association is not removed when moving the variable to a different function. It also does not differentiate a shallow copy of the value from taking a reference or a mutable reference. To put it shortly, it is an all-encompassing mapping between places and values.

To accommodate the possibility of linking the same value to multiple places, particularly when multiple memory locations hold an immutable reference to the value, it becomes necessary for the stored value to be a reference to the synchronization variable. To clarify, this introduces a second level of indirection. In order to facilitate the required cloning operations, we have opted to utilize `std::rc::Rc`, which is a smart pointer provided by the Rust standard library. The ownership of the referenced value (the synchronization variable) is shared and every time that

[36] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/mir_function/memory.rs

the value is cloned, an internal counter is incremented. When the count reaches zero, the value is freed [Klabnik and Nichols, 2023, Chap. 15.4].

The `Memory` utilizes a `std::collections::HashMap` data structure that establishes a mapping between `rustc_middle::mir::Place` instances and an enum with 5 variants corresponding to the 5 types mentioned previously that the translator tracks. 4 of these 5 variants enclose a `std::rc::Rc` reference to the synchronization variable. The aggregate case instead contains a vector of `Value`. This renders possible nesting aggregate values inside of each other, which is a critical requirement for supporting more complex programs with nested `structs`.

```rust
#[derive(Default)]
pub struct Memory<'tcx> {
    map: HashMap<Place<'tcx>, Value>,
}

/// ...

/// Possible values that can be stored in the `Memory`.
/// A place will be mapped to one of these.
#[derive(PartialEq, Clone)]
pub enum Value {
    Mutex(MutexRef),
    MutexGuard(MutexGuardRef),
    JoinHandle(ThreadRef),
    Condvar(CondvarRef),
    Aggregate(Vec<Value>),
}

/// ...

/// A mutex reference is just a shared pointer to the mutex.
pub type MutexRef = std::rc::Rc<Mutex>;

/// A mutex guard reference is just a shared pointer to the mutex guard.
pub type MutexGuardRef = std::rc::Rc<MutexGuard>;

/// A condvar reference is just a shared pointer to the condition variable.
pub type CondvarRef = std::rc::Rc<Condvar>;

/// A thread reference is just a shared pointer to the thread.
pub type ThreadRef = std::rc::Rc<Thread>;
```

Listing 4.6: A summary of the type definitions of the `Memory` implementation.

Using a hash map allows for efficient retrieval and management of the associated values during the translation process. The `Memory` also takes care of providing typedefs for the different references to synchronization variables. Listing 4.6 depicts an excerpt of the source file with the essential type definitions used in the implementation. Improvements to the current implementation are discussed in Sec. 6.5.

4.5.3. Intercepting assigments

The missing piece in the puzzle of the memory model is where to link the memory locations exactly. There are three separate places in the code in which this takes place.

On the one hand, the translator functions responsible for processing the methods of mutexes, condition variables, and threads create new synchronization variables that are linked to the return value of the corresponding method. This is where the lifetime of each synchronization variable starts. The specifics are expanded upon in Sec. 4.7.3 and 4.8.3.

On the other hand, the synchronization variable may be assigned in any other BB. For this reason, the translator incorporates a custom implementation of the method `visit_assign` to intercept every assignment in the MIR. Listing 4.7 shows precisely that all cases of copying, moving, or referencing the right-hand side (RHS) are handled by the same mechanism: The left-hand side (LHS) is linked to the right-hand side (RHS) if the type of the variable is a supported synchronization variable. The listing also shows how the compiler uses nested enums to model its data. Inside the variants of a right-hand side value (`rustc_middle::mir::Rvalue`), one can find operands (`rustc_middle::mir::Operand`). These operands also appear when passing function arguments.

The most peculiar case is the aggregate assignment. It materializes from assignments in Rust source code that create tuples, closures, or `structs`. It necessitates special handling as the value to be linked in memory must be assembled from the constituents of the aggregated value that are a synchronization variable. This implies that the `Memory` solely retains the portion of the aggregated value formed by the synchronization variables.

Tracking the assignments of synchronization variables at the moment they are returned from functions is another crucial mechanism. Fortunately, this can be accomplished by implementing a consistent check on all functions, regardless of whether they are modeled using the simple model (Fig. 3.2) or the function with cleanup model (Fig. 4.2). As a benefit, this uniform design readily supports `std::arc::Arc` without requiring any additional effort.

In every instance, the handling of assignments has no impact on the Petri net. No places or transitions are added when intercepting assignments.

Finally, some memory locations are passed to a new thread when calling `std::thread::spawn` and mapped again to the memory of the thread's function. The next section will demonstrate the method used to accomplish this.

```rust
fn visit_assign(
    &mut self,
    place: &rustc_middle::mir::Place<'tcx>,
    rvalue: &rustc_middle::mir::Rvalue<'tcx>,
    location: rustc_middle::mir::Location,
) {
    match rvalue {
        rustc_middle::mir::Rvalue::Use(
            rustc_middle::mir::Operand::Copy(rhs) | rustc_middle::mir::Operand::Move(rhs),
        )
        | rustc_middle::mir::Rvalue::Ref(_, _, rhs) => {
            let function = self.call_stack.peek_mut();
            link_if_sync_variable(place, rhs, &mut function.memory, function.def_id,
            ↪ self.tcx);
        }
        rustc_middle::mir::Rvalue::Aggregate(_, operands) => {
            let function = self.call_stack.peek_mut();
            handle_aggregate_assignment(
                place,
                &operands.raw,
                &mut function.memory,
                function.def_id,
                self.tcx,
            );
        }
        // No need to do anything for the other cases for now.
        _ => {}
    }

    self.super_assign(place, rvalue, location);
}
```

Listing 4.7: The custom implementation of `visit_assign` to track synchronization variables.

4.6. Multithreading

Multithreading support is a prerequisite for deadlock and missed signal detection. In order to support real-world programs where deadlocks or missed signals are possible in the first place, it becomes essential to support having several threads that share resources. First, the basics will be presented to later devise a PN model that captures the behavior of threads in Rust code.

4.6.1. Thread lifetime in Rust

The lifetime of a thread begins when it is started by invoking the `std::thread::spawn`[37] function. It receives a closure or function as an argument, representing the code that the new thread will execute concurrently with the other threads of the program. The spawned thread may start running immediately after it is spawned but there is no guarantee that it will do so.

Contrary to other programming languages like C, C++, or Java, Rust does not have the notion of a thread variable initialized previously to the start of the thread. Instead, the function `std::thread::spawn` returns a `std::thread::JoinHandle`, which is, as the name suggests, a handle to call `join` at the end of the thread's lifetime.

During its existence, a thread can independently execute its designated code and perform various operations concurrently with other threads. It can access shared resources and communicate with other threads through synchronization mechanisms like mutexes, condition variables, channels, or atomic operations. This enables concurrent processing and parallelism in Rust programs.

To ensure proper coordination between threads, Rust provides a mechanism to join threads. The `std::thread::JoinHandle::join`[38] method allows the main thread or another thread to wait for the completion of a different thread. By calling join on a join handle, the calling thread blocks until the spawned thread finishes its execution. Once a thread completes its execution and is joined by another thread, its lifetime ends, and the corresponding system resources are released. Otherwise, threads that were not joined correctly may potentially leak resources.

If the join handle is dropped, the thread may no longer be joined and it implicitly becomes *detached*. A detached thread refers to a thread without a valid join handle. It will continue its execution independently until it completes or the program terminates. They are useful in scenarios where the spawning thread does not need to wait for the thread to complete its task. For example, in long-running background tasks or when the main thread terminates independently of the detached thread's progress. However, it's important to stress that the execution of detached threads may continue *even* after the main thread exited.

4.6.2. Petri net model for a thread

To incorporate additional threads into the PN model, a distinct subnet is appended to the main net to represent each thread. This subnet encapsulates the execution path of the newly spawned thread and operates as an isolated context. It establishes precise interfaces that connect back to the main net. The closure provided to the `spawn` function, being a MIR function, can invoke other functions that in turn require translation. Therefore, processing a thread's function follows a similar approach to the usual translation logic.

[37] https://doc.rust-lang.org/std/thread/fn.spawn.html
[38] https://doc.rust-lang.org/std/thread/struct.JoinHandle.html#method.join

The concurrency aspect of the execution of the new thread is modeled by the generation of a new token at the transition that represents the call to `spawn`. This token can be interpreted, in the same manner as the token in `PROGRAM_START`, as the instruction counter of the new thread. Essentially, the spawn operation constitutes a "fork" in the token flow: One token enters the transition and two tokens exit from it. The first proceeds along the main thread's path to execute the subsequent statement, while the second is directed to the first BB of the function passed to the thread.

Each thread identified in the source code possesses designated start and end places labeled `THREAD_<index>_START` and `THREAD_<index>_END`, respectively. The index is mandatory to preserve the label uniqueness property across the entire program. It should be emphasized that this mimics the basic places for the program detailed in Sec. 4.1.1.

Threads lack a separate panic place as invoking `panic!` inside a thread only terminates that specific thread's execution. We are not interested in differentiating between thread end states; the main requirement is to determine whether a thread has finished or not. A single end place for both cases suffices here.

The joining behavior serves as the inverse operation of the spawn. The transition corresponding to the `join` call consumes two tokens but generates only one token. As a result, the waiting condition is modeled straightforwardly: The main thread can continue, i.e, the `join` transition can fire, if and only if the thread to be joined has finished execution, reaching its respective `THREAD_END` place.

To recapitulate, the thread is translated to a separate subnet that interfaces with the main net only at three places:

- The `spawn` transition where the thread starts.
- The (optional) `join` transition where the join handle is utilized.
- The connections due to synchronization variables, analyzed later in the dedicated sections of this chapter.

4.6.3. A practical example

Observe Listing 4.8 and its corresponding PN model in Fig. 4.6. This is one of the test programs found in the repository. Note the "fork" at the spawn transition described in the previous subsection. The left branch is the thread, while the right branch is the main thread. It is clear that the paths split at the `spawn` and merge at the `join`. Notice also that there is no separate panic place for the thread, indicating that a failure in one thread does not impact the other threads.

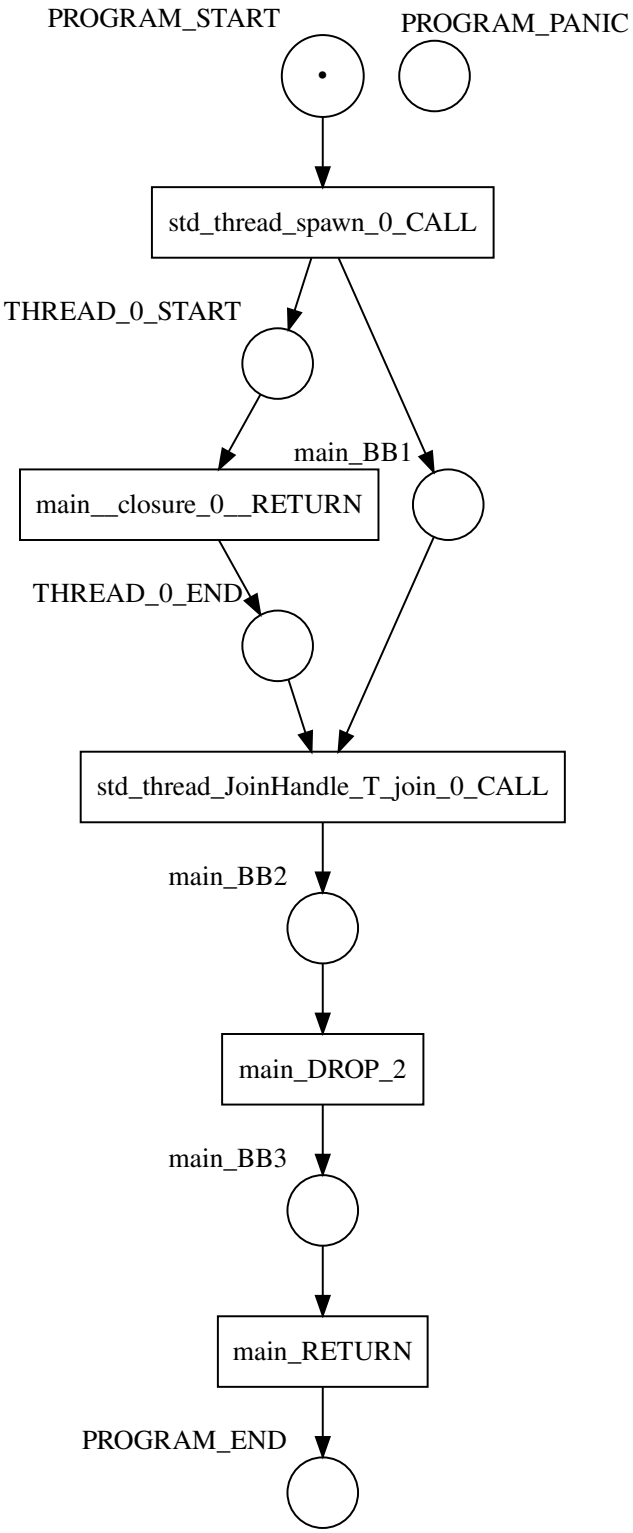

Figura 4.6: The Petri net model for the program in Listing 4.8.

```
fn main() {
  let thread_join_handle = std::thread::spawn(move || {
      // some work here
  });
  // some work here
  let _res = thread_join_handle.join();
}
```

Listing 4.8: A basic program with two threads to demonstrate multithreading support.

4.6.4. Algorithms for thread translation

To close this section, we will briefly describe the algorithms used for translating threads. Initially, it is worth mentioning that since the translation is carried out by a single thread (the tool does not support multiple threads translating the source code), a decision has to be made concerning when to translate spawned threads:

- Immediate translation: Translate the thread as soon as it is encountered. The translator "switches" to the spawned thread.
- Delayed translation: Store all the relevant information about the new thread and translate it after the main thread.

The current solution takes the latter approach.

When a call to `std::thread::spawn` is encountered:

1. Translate the function call using the model seen in Fig. 4.2.
2. Retrieve the first argument passed to the function: An aggregate value that holds the variables captured by the closure and the function to be executed by the thread.
3. Extract the ID of the function to be executed by the thread.
4. Extract the values captured by the closure.
5. Create a new `Thread`[39] to store the information required for the delayed translation.
6. Link the return value of `std::thread::spawn`, the new join handle, to the `Thread`.
7. Push the thread to a queue of detected threads in the `Translator`.

When a call to `std::thread::JoinHandle::join` is encountered:

1. Translate the function call using the model seen in Fig. 3.2. Ignore the cleanup place since we must force the PN to "wait" for the thread to exit. This is equivalent to assuming that the `join` function never fails.

[39] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/sync/thread.rs

2. Retrieve the first argument passed to the function: The join handle. The memory location is linked to the corresponding thread thanks to the assignment interception explained in Sec. 4.5.3.
3. Set the join transition of the underlying `Thread` behind the join handle.

When the main thread finishes translating, that is, when the `main` function has already been processed, the `Translator` enters a loop to translate the threads discovered so far in order.

1. Create a new start and end place for the thread.
2. Connect the spawn transition to the start place.
3. If a join transition was found, connect the end place to it.
4. Replace the place `PROGRAM_PANIC` with the place `THREAD_<index>_END` to translate terminators like `Unwind` correctly (Sec. 4.4.3).
5. Push the thread function to the call stack.
6. Move the synchronization variables to the memory of the thread function, i.e., map the aggregate value and its fields to the memory of the thread function.
7. Translate the top of the call stack.

As anticipated before, the algorithm exhibits resemblances to the overall procedure for function calls outlined in Sec. 4.2. Lastly, the implementation is capable of handling programs where threads spawn their own threads in a nested manner. The threads are simply added to the queue and as the loop advances, the nested threads are translated too.

4.7. Mutex (`std::sync::Mutex`)

A mutex, short for mutual exclusion, is a synchronization mechanism used to control access to a shared resource in a concurrent program. It allows multiple threads to access the shared resource in a mutually exclusive manner, ensuring that only one thread can access the resource at a time.

In this section, the PN model for a mutex in Rust is explained, then a practical example is presented to ease comprehension and finally the algorithms used for the translation of mutex functions are outlined.

4.7.1. Petri net model

In Rust, a mutex is created by wrapping the shared data in a `Mutex<T>` type, where `T` is the type of the shared resource. The `std::sync::Mutex` type exposes a method named `lock` to acquire the

lock on the shared resource. If the mutex is currently unlocked, the thread successfully acquires the lock and can proceed with accessing the resource. If the mutex is already locked by another thread, the thread attempting to acquire the lock will be blocked until the lock becomes available. The `lock` method returns a mutex guard (`std::sync::MutexGuard`) that grants exclusive access to the resource until it is dropped.

Contrary to the `unlock` semantics present in C or C++, the mutex included by the Rust standard library is unlocked implicitly, i.e., without calling a function. The mutex implements Resource Acquisition Is Initialization (RAII) and releases the lock automatically when it goes out of scope, preventing deadlocks. Alternatively, dropping a local variable of type `std::sync::MutexGuard` is equivalent to unlocking the corresponding mutex.

A mutex can be modeled in PN as a single place that represents the state of the mutex, indicating whether it is locked or unlocked. The place is labeled to reflect its purpose as a mutex. Besides, the place is marked with a token initially to signify that the mutex starts in the unlocked state.

Transitions that lock the mutex consume the token from the mutex place. If the token is absent, the transition may not fire. The mutex must be in the unlocked state to enable the locking transition, which is the desired behavior.

Transitions that unlock the mutex produce a token at the mutex place. The transition can fire as long as the program reached that point in the execution. After the transition fires, the mutex place holds again a token that can be consumed by a locking transition. Two types of transitions may unlock the mutex:

1. A `Drop` terminator (Sec. 4.4.3) when the dropped place is of type `std::sync::MutexGuard`.
2. The transition for a call to `std::mem::Drop`, which frees the memory occupied by the value passed in explicitly.

By connecting the mutex place to the locking and unlocking transitions using input and output arcs, we establish the relationship between the mutex state and the actions that manipulate it. This modeling approach allows for the representation of the mutex's behavior in a PN and facilitates the analysis of its interactions with other parts of the system.

The PN model presented here is well-known in the literature and has been applied successfully in other tools. It can be found among others in [Kavi et al., 2002, Moshtaghi, 2001, Meyer, 2020, Zhang and Liua, 2022].

4.7.2. A practical example

Consider the PN model shown in Fig. 4.7 corresponding to the program in Listing 4.4. The MIR is depicted in 4.5. This test program is one of the examples included in the repository.

Observe that there are two locking transitions mapping the two calls to `lock` in the source code. The indexing system mirrors the order of their appearance in the program, which justifies the labels `std_sync_Mutex_T_lock_0_CALL` and `std_sync_Mutex_T_lock_1_CALL`. Both have an incoming arc from the mutex place `MUTEX_0`.

As explained before, `Drop` terminators may unlock a mutex. No matter if they fail or not (the error case includes the suffix `_UNWIND`), an outgoing arc flows back to the mutex place to replenish the token.

One should take note that there are more incoming arcs to the mutex place than outgoing arcs, which highlights the importance of following the mutex guards throughout the MIR using the strategy explained in Sec. 4.5.3.

4.7.3. Algorithms for mutex translation

Concluding this section, we will provide a brief overview of the algorithms employed in the translation of mutex functions.

When a call to `std::sync::Mutex::new` is encountered:

1. Translate the function call using the model seen in Fig. 4.2.
2. Create a new `Mutex`[40] structure with an index to identify it unequivocally across the PN.
3. Link the return value of `std::sync::Mutex::new`, the new mutex, to the `Mutex` structure.

When a call to `std::sync::Mutex::lock` is encountered:

1. Translate the function call using the model seen in Fig. 3.2. Ignore the cleanup place since we must force the PN to "wait" for the mutex place to be marked. This is equivalent to assuming that the `lock` function never fails.
2. Retrieve the `self` reference to the mutex on which the function is called.
3. Add an arc from the underlying mutex place to the transition representing the function call.
4. Create a new `MutexGuard` with a reference to the `Mutex`.
5. Link the return value of `std::sync::Mutex::lock`, the new mutex guard, to the `MutexGuard` structure.

When a call to `std::mem::drop` is encountered:

1. Translate the function call using the model seen in Fig. 4.2.
2. Extract the variable passed into the function.

[40] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/sync/mutex.rs

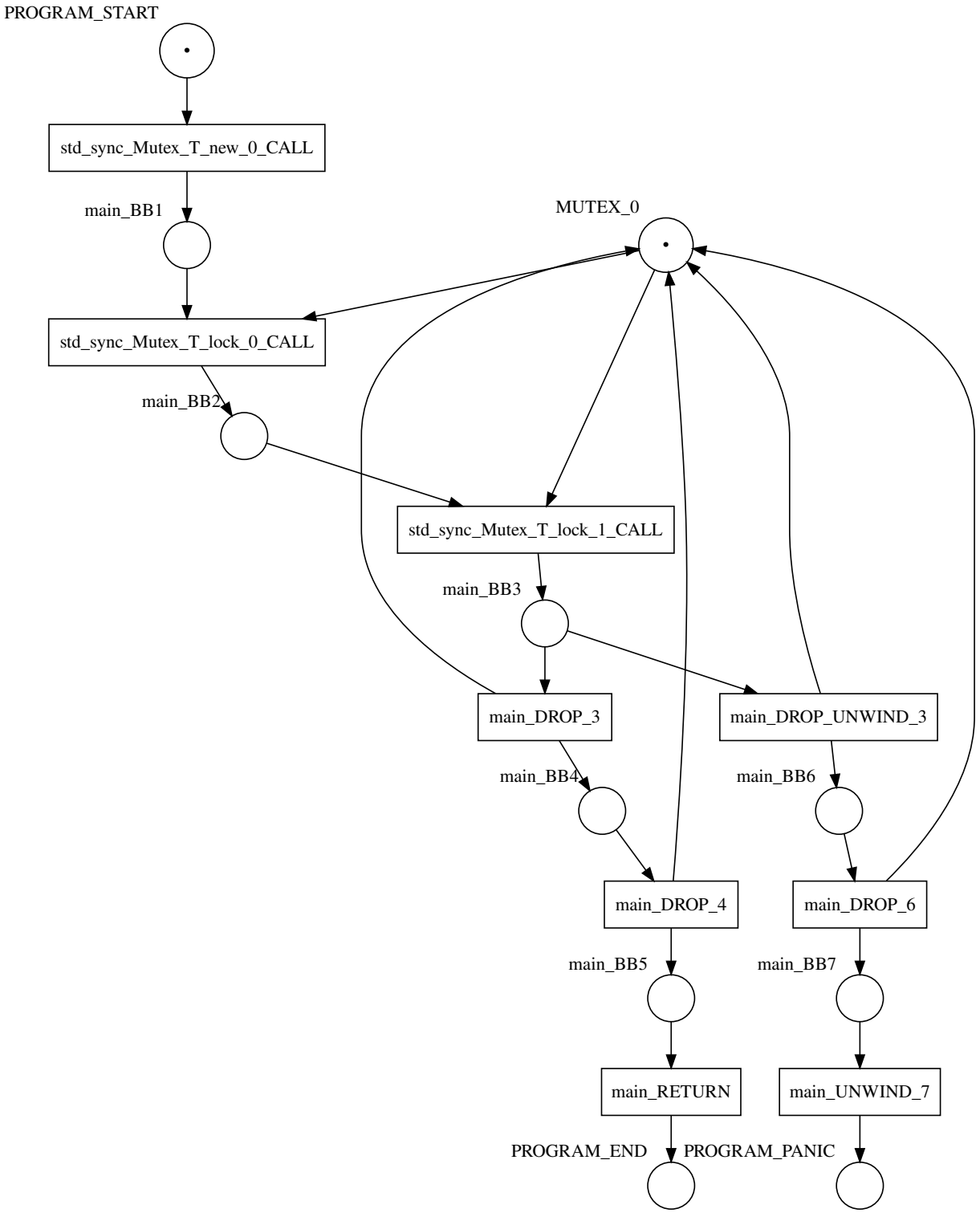

Figura 4.7: The Petri net model for the program in Listing 4.4.

3. If the variable is linked to a mutex guard, add an arc from the transition of the function call to the mutex place.
4. If a cleanup place was provided, add an unlock arc from the cleanup transition to the mutex place too.

When a terminator of kind `rustc_middle::mir::TerminatorKind::Drop` is encountered:

1. If the variable to be dropped is linked to a mutex guard, add an arc from the transition of the function call to the mutex place.
2. If a cleanup place was supplied, add an unlock arc from the drop unwind transition to the mutex place too.

In the upcoming section, we will delve into the necessary adjustments of these algorithms to establish a unified model for condition variables, which is essential for detecting missed signals. Given that these modifications are better comprehended within the framework of condition variables, we will elucidate them in that specific context.

4.8. Condition variable (`std::sync::Condvar`)

A condition variable is a synchronization primitive used in concurrent programming to enable threads to wait for a certain condition before proceeding with their execution. Threads wait until they are notified by another thread that the desired condition has been met.

Condition variables are typically associated with a mutex, which ensures exclusive access to the shared data that the condition depends on. When a thread waits on a condition variable, it releases the associated mutex, allowing other threads to make progress. When the condition becomes true or some event occurs, a notifying thread signals the condition variable, allowing one or more waiting threads to resume their execution.

The semantics of condition variables, as well as examples in pseudocode, were introduced in Sec. 1.5. The understanding of the precise behavior of condition variables in all circumstances is a prerequisite for this section.

This section provides an elaborate explanation of the PN model used to represent condition variables from the Rust standard library. It is followed by a practical example that aims to enhance the clarity of the concepts. Finally, the algorithms for the translation of condition variable functions are outlined.

4.8.1. Petri net model

In this particular case, the PN model must be examined carefully, as it involves not only the condition variable itself but also the variable that holds the condition on which the blocked

thread is waiting *and* the mutex that synchronizes access to that condition.

This interaction can be extremely complex in general. For instance, the same condition variable could be used to signal an arbitrary number of distinct conditions. Accordingly, different mutexes may be passed as an argument to the `wait` call. Furthermore, an arbitrary number of threads may block on a condition variable and Rust supports the broadcast operation to awaken all waiting threads at once through the method `notify_all`[41] (see Sec. 1.5). Most importantly, the condition itself could be of any type and could take a long sequence of values during the execution, depending on which the waiting threads could act in diverse ways for every scenario.

For the above reasons, it is unavoidable to make assumptions regarding the supported use cases of condition variables in order to reduce the complexity of the task. Embracing and dealing with every possibility is beyond the scope of this thesis.

Assumptions

1. *Single call*: There is only one call to `wait` per condition variable, i.e., `condvar.wait()` appears in a single place in the source code for a given `condvar`. For example, it can be inside of a loop but it cannot be in two different functions.
2. *Single-element queue*: There is at most one waiting thread per condition variable.
3. *Boolean condition*: The condition is a boolean flag. It is either set or not set. Waiting on a condition that may take 3 or more values is *not* supported by this model.
4. *Mandatory set-condition / No "false notify"*: If a thread locks the mutex and accesses the shared condition mutably, then it always sets it to a different value. In simpler terms, threads that look at the value, do not change it and immediately notify the condition variable are *not* supported.
5. *Broadcast exclusion*: The method `std::sync::Condvar::notify_all` is out of scope.

Support for multiple calls to `wait` and multiple waiting threads could be implemented but a considerable implementation effort is required. Therefore, assumptions 1 and 2 can be overcome with the proposed model.

Supporting non-boolean conditions and detecting which value is set necessitates a thorough reconsideration of the modeling approach for representing concrete data values in simple Petri nets. Consequently, assumptions 3 and 4 are particularly challenging and could be the subject of future research into higher-lever models. See Sec. 6.6 for some thoughts to this effect.

[41] https://doc.rust-lang.org/std/sync/struct.Condvar.html#method.notify_all

Analysis of the proposed model

Fig. 4.8 depicts the PN model used in the implementation. The same diagram in DOT, PNG, and SVG format can be found in the repository as documentation.

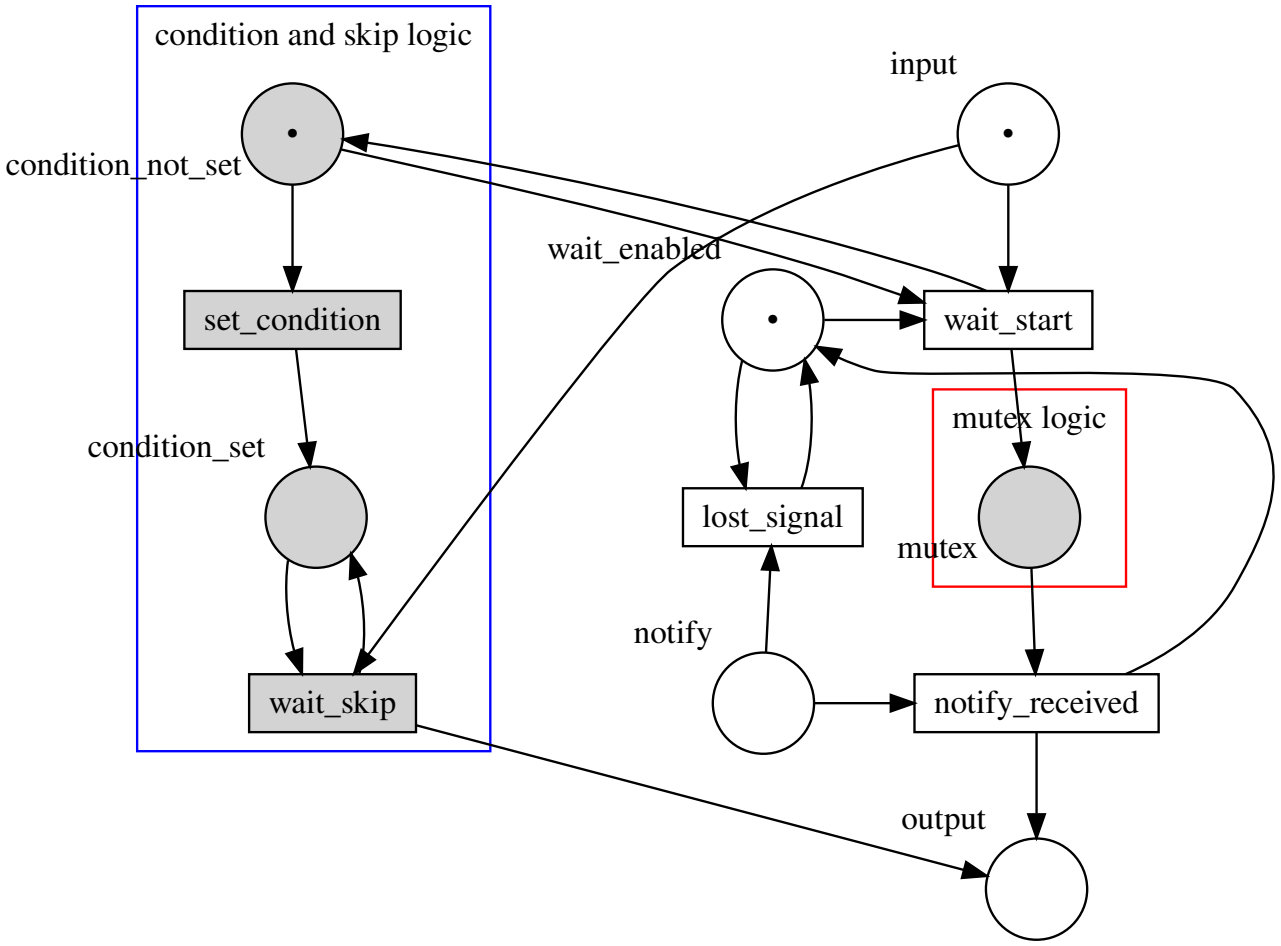

Figura 4.8: The Petri net model for condition variables.

The input places are:

- `input`: The start place of the wait function. The model supports the methods from the standard library `std::sync::Condvar::wait` and its variation `std::sync::Condvar::wait_while`.
- `condition_not_set`: The place is marked when the condition is `false`.
- `condition_set`: The place is marked when the condition is `true`.
- `notify`: The place where the notifying thread places a token to wake up the waiting thread.

The `output` place is the end place of the function call to `wait` or `wait_while`. The execution of the thread continues from there.

Two possible ways exist to go from `input` to `output`, represented by two transitions:

- `wait_start`: This is the "common case", the thread blocks and waits for the signal.
- `wait_skip`: This is the alternative path that the token takes when the condition was already set. The thread does not wait, instead, it skips the wait and reaches `output` in a single jump.

It is essential to notice that the greyed-out part on the left of Fig. 4.8 controls which transition is enabled and which transition is disabled. As soon as a token is set in `condition_set`, `wait_start` is disabled. Before that, the opposite is true: `wait_start` may fire but `wait_skip` may not.

Note the arcs between `condition_not_set` and `wait_start`. The token is regenerated every time that `wait_start` fires. The same is true for `condition_set` and `wait_skip`. These arcs restore the condition places to their previous state. Modeling more than 2 values as a PN would entail a more convoluted net, Besides, it would be impossible to know at compile time, how many possible values the condition takes to generate the correct number of places. This limitation is the justification for Assumptions 3 and 4.

Now focus on the right side of Fig. 4.8. In the middle of the condition variable, we find the place for the mutex. As expected, it is unlocked when the wait starts and it is locked when the notify is received.

The place labeled `wait_enabled` plays an important role. On the one hand, it consumes the token from `notify` if `wait_start` was not fired. This is the archetypical missed signal case that we would like to detect. On the other hand, the token in `wait_enabled` is consumed when `wait_start` fires, This prevents the condition variable from "accepting" other threads (Assumption 2) and preserves the token in `notify`, ensuring that the missed signal cannot occur.

Finally, the `notify_received` transition combines the requisites for the thread to leave the wait: The mutex must be unlocked and `notify_one` was called. To restore the initial state of the condition variable, it regenerates the token in `wait_enabled`.

Global translation requirements of the Petri net model

A fundamental challenge that surfaces during the implementation of the model in Fig. 4.8 is that connections across the blue frontier in the diagram cannot be established in general when processing the call to `wait`. We will analyze the problem and explain how the solution copes with it.

The transition where the condition is set, named `set_condition` on the diagram, is the next candidate. In the current implementation, the transition selected to fulfill this role is the call to `std::ops::DerefMut::deref_mut` when a mutex or mutex guard is being dereferenced.

Consider Listing 4.9, yet another test program from the repository. In line 9, the mutex guard is dereferenced to write the value `true` to it, setting the condition for the condition variable. In

the MIR, this maps to a call to `std::ops::DerefMut::deref_mut`. Although it is not the exact place where the value is written (which would actually be a statement in BB), it is close enough and it satisfies our needs.

```
fn main() {
      let pair = std::sync::Arc::new((std::sync::Mutex::new(false),
      ↪  std::sync::Condvar::new()));
      let pair2 = std::sync::Arc::clone(&pair);

      // Inside of our lock, spawn a new thread, and then wait for it to start.
      std::thread::spawn(move || {
          let (lock, cvar) = &*pair2;
          let mut started = lock.lock().unwrap();
          *started = true;
          // We notify the condvar that the value has changed.
          cvar.notify_one();
      });

      // Wait for the thread to start up.
      let (lock, cvar) = &*pair;
      let mut started = lock.lock().unwrap();
      while !*started {
          started = cvar.wait(started).unwrap();
      }
}
```

Listing 4.9: A program that requires global Petri net information to be translated.

Unfortunately, the connections to the condition variable cannot be established when processing the call to `deref_mut` either. The reason is that there is no guarantee that the condition variable was already seen. It could still be ahead in the translation path. In Listing 4.9, the main thread is translated first, so the condition variable is discovered first. But if the roles of the threads are swapped, then the translation cannot be performed.

We reach thus an unpleasing conclusion. In order to connect the model of the condition variable to the places that model the condition and the transitions where it is set, we need the whole PN, i.e., we need *global* information to translate the synchronization primitive effectively.

As a result, it is unavoidable to incorporate some sort of postprocessing step to the translation. The tasks must also be performed in a specific order. The mutex must have been discovered first. Later it may be linked to a condition variable if such a condition variable is found (the source code may as well not use any). Hence, it is advisable to introduce a notion of "priority" to the postprocessing tasks.

The `Translator` relies on a `std::collections::BinaryHeap` to implement a priority queue of `PostprocessingTask`[42]. The tasks are returned by the methods that translate synchronization primitives if needed. After all the threads are translated, the `Translator` addresses the postprocessing tasks. By completing them according to their priority, we guarantee that the information is available in the required order.

Table of possible inputs and expected outputs

As a complement to the explanation in the previous subsections, here is a table summarizing the expected output for a given input. The reader may compare Fig. 4.8 to verify that the model produces the correct output for each scenario.

Row #	Input			Output
	`condition_set`	`wait_enabled`	`notify`	*where the initial token at* ***`input`*** *ends*
R1	False	False	False	`waiting` (waiting for a `notify`)
R2	False	False	True	`output` (correct wait end condition)
R3	False	True	False	`input` (initial state)
R4	False	True	True	*lost signal (transient state, goes to R1)*
R5	True	False	False	`waiting` (condition set, needs `notify`)
R6	True	False	True	`waiting` (correct wait end condition)
R7	True	True	False	`output` (skip the wait)
R8	True	True	True	`output` (skip the wait, with lost signal)

Cuadro 4.1: A summary of the possible states of the Petri net model for condition variables.

4.8.2. A practical example

Due to the size constraints of the resulting PN, we are compelled to select a small-scale program for demonstration purposes. It would be unfeasible to embed within a single page the complete PN of a realistic program with condition variables and multiple threads. For more comprehensive examples, readers are encouraged to explore the repository, which contains a collection of more intricate programs included as part of the integration tests.

Despite the space limitations, the example in Listing 4.10 comprises the core elements of the model presented before. The complete PN can be seen in Fig. 4.9.

Observe the following sequence of transitions:

1. The mutex is locked in `std_sync_Mutex_T_lock_0_CALL`.

[42] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/function.rs#L92

2. std_sync_Condvar_notify_one_0_CALL sets a token in CONDVAR_0_NOTIFY.
3. The token flow continues to main_BB5 right before CONDVAR_0_WAIT_START.

It should be emphasized that no deadlock arises if the transition CONDVAR_0_WAIT_START fires *before* CONDVAR_0_LOST_SIGNAL. In short, a conflict exists between CONDVAR_0_WAIT_START and CONDVAR_0_LOST_SIGNAL for the token in CONDVAR_0_NOTIFY. Nevertheless, the model checker verifies *all* possible firings and it will uncover the missed signal case without difficulties.

Another noteworthy observation is that this program illustrates the effect that the cleanup paths would have on missed signal detection. If there were a second transition at the same level as CONDVAR_0_WAIT_START or std_sync_Condvar_notify_one_0_CALL, the token could "escape" to the PROGRAM_PANIC place and the deadlock would remain undetected.

It is indispensable for the translation to "force" the Petri net to stay blocked and not open alternative paths that could be used by the model checker to come to the conclusion that the PN never deadlocks.

```
1 fn main() {
2       let mutex = std::sync::Mutex::new(false);
3       let cvar = std::sync::Condvar::new();
4       let mutex_guard = mutex.lock().unwrap();
5       cvar.notify_one();
6       let _result = cvar.wait(mutex_guard);
7 }
```

Listing 4.10: A basic program to showcase condition variable translation.

4.8.3. Algorithms for condition variable translation

To wrap up this section, we will present a concise summary of the algorithms utilized in the translation of condition variables. The additions required to the mutex algorithms are included afterward as well.

When a call to std::sync::Condvar::new is encountered:

1. Translate the function call using the model seen in Fig. 4.2.
2. Create a new Condvar[43] structure with an index to identify it unequivocally across the PN.
3. Link the return value of std::sync::Condvar::new, the new condition variable, to the Condvar structure.

[43] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/translator/sync/condvar.rs

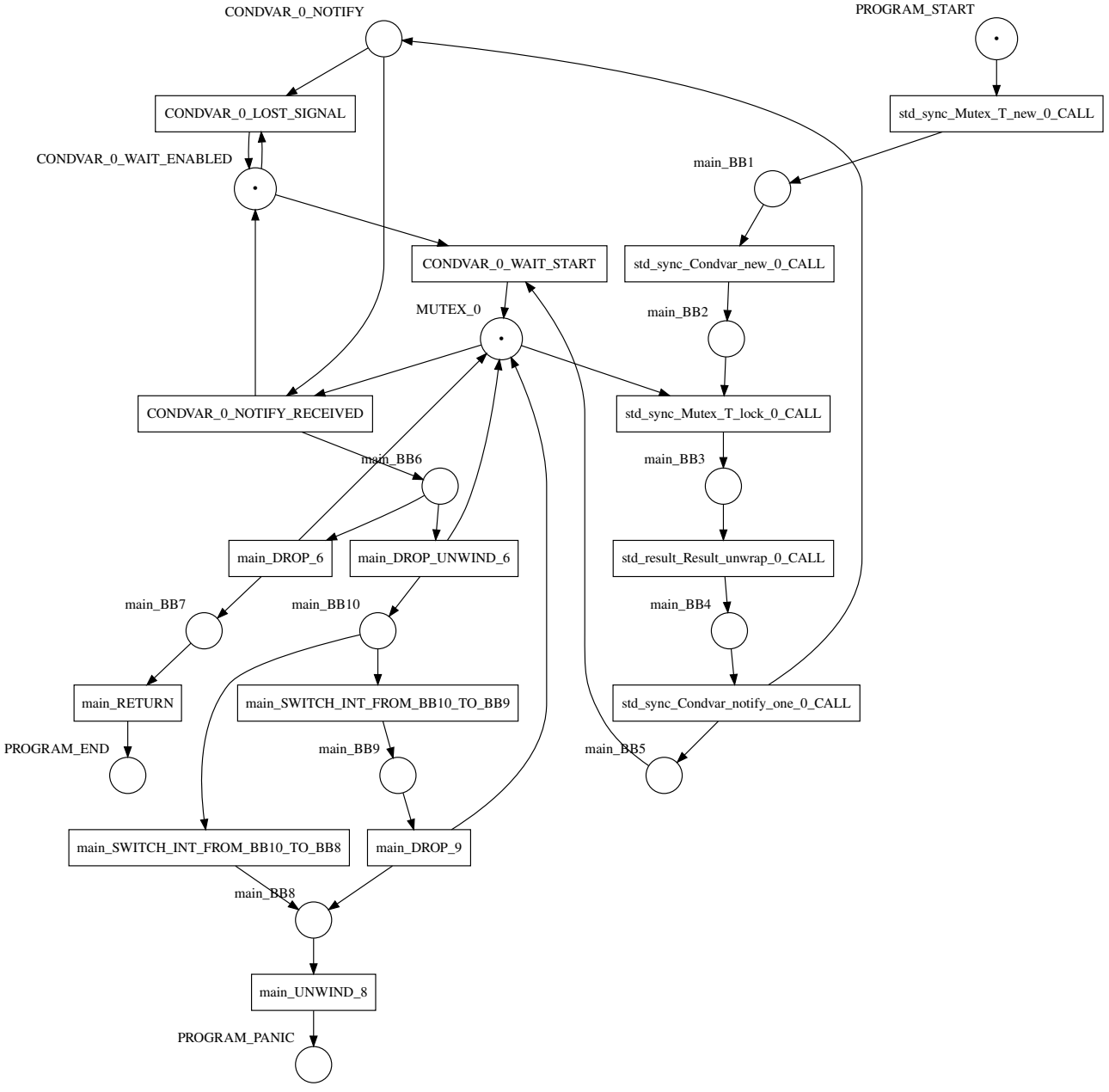

Figura 4.9: The Petri net model for the program in Listing 4.10.

When a call to `std::sync::Condvar::notify_one` is encountered:

1. Translate the function call using the model seen in Fig. 3.2. Ignore the cleanup place because, otherwise, any call may fail, which amounts to the notify operation not being present in the program, leading to a false lost signal. This is equivalent to assuming that the `notify_one` function never fails.
2. Retrieve the `self` reference to the condition variable on which the function is called.
3. Add an arc from the transition representing the function call to the `notify` place of the underlying condition variable.

When a call to `std::sync::Condvar::wait` or `std::sync::Condvar::wait_while` is encountered:

1. Ignore the cleanup place because, otherwise, any call may fail, which amounts to the wait operation not being present in the program, leading to an incorrect result. We must force the PN to "wait" for the notifying signal to be sent. This is equivalent to assuming that the `wait` or `wait_while` function never fails.
2. Retrieve the `self` reference to the condition variable on which the function is called.
3. Extract the mutex guard passed into the function.
4. If the condition variable was already connected to a function call, then the translation fails. This enforces Assumptions 1 and 2 seen at the beginning of the section.
5. Otherwise, connect the start and end places to the `wait_start` and `notify_received` transitions respectively.
6. Link the return value, the same mutex guard that was passed as an argument, to the `MutexGuard` structure.
7. Notify the translator that the mutex received must be linked to this `Condvar`. For this purpose, use the enum variant `PostprocessingTask::LinkMutexToCondvar`. This task will be processed after translating all the threads.

When all the threads finished translating, that is, when the queue of threads to process is empty, the `Translator` enters a loop to complete the postprocessing tasks by priority order:

1. Create at the beginning of the loop an empty vector of mutex references.
2. Pop from the `std::collections::BinaryHeap` the task with the lowest priority. This will be by design a `PostprocessingTask::NewMutex`. Add the mutex reference to the vector.
3. After processing all the lower priority tasks, the `Translator` has references to all the mutexes in the code. Continue popping tasks from the priority queue.
4. Eventually, a `PostprocessingTask::LinkMutexToCondvar` is extracted. Link each mutex to the condition variable, which creates the places `condition_set` and `condition_not_set` for the condition. It also connects the `deref_mut` transitions to these places to set the

condition. Lastly, it connects the condition places to the condition variable transitions to disable the `wait`.

Modifications to the mutex algorithms

As stated before, the mutex algorithms require some additions to successfully perform missed signal detection.

Add the following to the handler of the `std::sync::Mutex::new` function:

1. Notify the translator that a new mutex has been created. For this purpose, use the enum variant `PostprocessingTask::NewMutex`. This task will be processed after translating all the threads.

When a call to `std::result::Result::<T, E>::unwrap` is encountered:

1. Check that the `self` reference is a mutex or a mutex guard.
2. Translate the function call using the model seen in Fig. 3.2. Ignore the cleanup place because, otherwise, any call may fail, as if the mutex lock operation were not present in the program, leading to a false lost signal. This is equivalent to assuming that the `unwrap` function never fails when applied to a variable linked to a mutex or a mutex guard.

When a call to `std::ops::Deref::deref` or `std::ops::DerefMut::deref_mut` is encountered:

1. Check that the `self` reference is a mutex or a mutex guard.
2. Translate the function call using the model seen in Fig. 3.2. Ignore the cleanup place because, otherwise, any call may fail, as if the mutex lock operation were not present in the program, leading to a false lost signal. This is equivalent to assuming that the `deref` and `deref_mut` functions never fail when dereferencing a variable linked to a mutex or a mutex guard.
3. If the value is being dereferenced mutably (`deref_mut`), extract the first argument passed to the function: The mutex or mutex guard. Add the `deref_mut` transition to the mutex to set the condition for a condition variable in the postprocessing step.
4. Otherwise do nothing. The immutable case does not need to be added to the mutex.

It should now be clear to the reader that the algorithms for missed signal detection are fundamentally of higher complexity and ought to handle more border cases than those for detecting simple deadlocks caused by incorrect usage of mutexes or calling `join` on threads that never terminate.

It is worth mentioning that some border cases arise due to the inclusion of the cleanup logic from the MIR in the PN model. If the implementation instead skipped this, under the hypothesis that Rust standard library functions may never panic, then the algorithms would become simpler. Sec. 6.2 is concerned with the question of not modeling the cleanup paths.

Capítulo 5

Probando la implementación

The inclusion of a dedicated chapter on testing in the thesis underscores the significance of this indispensable aspect of the development process. Tests play a fundamental role in ensuring the reliability and correctness of the software implementation. A comprehensive test suite has been developed to cover the extensive functionality and behavior of the translator and the PN library.

The tests encompass multiple levels that will be elucidated in the subsequent sections. At the lowest level, unit tests are conducted to verify the correctness of the data structures employed within the translator and the PN library. These tests target individual components, thoroughly examining their functionality in isolation.

In addition to unit tests, a suite of integration tests has been constructed incrementally to evaluate the translator's adherence to expected behavior. These tests consist of test programs that simulate simple scenarios where the resulting file output is compared against the expected results. This testing methodology helps to uncover any regressions in the compiler and confirms that the translator functions reliably in the supported use cases.

Furthermore, we incorporated a description of how to generate the MIR and visualize the result of the translation to aid in the debugging process. The tooling enables exposing the internal details in an accessible and understandable manner.

Later in this chapter, the usage of the model checker LoLA and its integration within the translator is explained. The model checker provides more features than the minimal set that was integrated into the translator to answer the deadlock detection problem. Hence, it is beneficial to explore which features the model checker provides for debugging the PN translation.

Finally, the capabilities of the tool are demonstrated by means of two test programs that model classical problems in concurrent programming.

5.1. Unit tests

The unit tests form the base of the test suite. The PN library and the data structures used in the translator rely on them extensively. By testing the underlying data structures meticulously, potential issues can be identified and resolved early in the development cycle before continuing work on the higher-level components of the translator.

5.1.1. Petri net library

The PN library `netcrab` uses unit tests to verify that adding places, transitions, and arcs to a net behaves as expected. The translator performs these operations often so it is important to verify them. The iterators on which the export formats are built are tested as well.

On the other hand, each one of the three export formats (DOT, PNML and LoLA) comes with unit tests to check that the output is generated correctly for the following simple cases:

- Empty net.
- A PN with 5 places and 0 transitions.
- A PN with 0 places and 5 transitions.
- A PN with 5 places marked with different numbers of tokens.
- A PN with a chain topology.
- A PN with 1 place and 1 transition connected in a loop.

These tests helped to troubleshoot bugs related to the LoLA format. For concrete examples, see this commit[1] or this other commit[2].

5.1.2. Stack

The simple `Stack`[3] data structure is employed to implement the call stack in the `Translator`, as seen in Sec. 4.2. Unit tests demonstrate the supported methods and check some simple use cases.

[1] https://github.com/hlisdero/netcrab/commit/5745e0da5d27bd709ef479f45a6d2e75974d3745
[2] https://github.com/hlisdero/netcrab/commit/dbce3f8999ece32e6731527c303a7b59858991f9
[3] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/data_structures/stack.rs

5.1.3. Hash map counter

Analogous to the stack, the `HashMapCounter`[4] contains some unit tests to verify that the methods work as intended. This data structure forms the basis for the function counter in the `Translator` that keeps track of how many times each function was called to generate a unique incremental index for the transition labels.

5.2. Integration tests

We will now examine the integration tests, which serve as the backbone for the translator's testing framework. Two types of tests exist currently:

- Translation tests.
- Deadlock detection tests.

The testing process has proven invaluable throughout the development of the translator, enabling early detection of bugs and regressions in *rustc*. By relying on the previous tests and building features incrementally, the implementation progressed with confidence and a firm step forward. The testing capabilities offered by Rust, including its support for unit tests and integration tests, have been instrumental in ensuring the quality and reliability of the translator.

5.2.1. Translation tests

In translation tests, a given program is processed *without* performing the deadlock analysis. As a result, three text files containing the model in DOT, PNML, and LoLA formats are generated. These files are then compared to the expected output, which is stored in the repository and serves as documentation as well.

The expected output was verified manually using the tools presented in Sec. 5.3. It was committed to the repository when the translator was first able to pass the test. If a regression in *rustc* occurs, then the expected output files are updated accordingly. This has happened some times in the past. See for instance this commit[5] or this one[6].

[4] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/data_structures/hash_map_counter.rs

[5] https://github.com/hlisdero/cargo-check-deadlock/commit/881a3873a3b060e70bc727f670f9426d14327fa2

[6] https://github.com/hlisdero/cargo-check-deadlock/commit/b032fa3cc13e631950a802dcd3f755c548afde86

5.2.2. Deadlock detection tests

Deadlock detection tests are closer to an end-to-end test of the translator. They generate the file in LoLA format for the test program and instruct the translator to perform the deadlock analysis. The output is then contrasted to the known behavior of the test program, i.e., it deadlocks or it does not deadlock. If LoLA produces an incorrect result, then the test fails. In such a case, the PN model should be analyzed to find the source of the error. See Sec. 5.3.3 for details on how to approach this.

Observe Listing 5.1, which contains the contents of the `.lola` file for the program depicted in Listing 4.4. This is the file format that the model checker requires. It is relatively simpler than PNML, which is XML-based.

Due to the considerable length of the output, it is the sole instance of the LoLA format in this thesis. It is included here for completeness. The repository contains several other examples, all of which are used in the integration tests.

```
1  PLACE
2      MUTEX_0,
3      PROGRAM_END,
4      PROGRAM_PANIC,
5      PROGRAM_START,
6      main_BB1,
7      main_BB2,
8      main_BB3,
9      main_BB4,
10     main_BB5,
11     main_BB6,
12     main_BB7;
13
14 MARKING
15     MUTEX_0 : 1,
16     PROGRAM_END : 0,
17     PROGRAM_PANIC : 0,
18     PROGRAM_START : 1,
19     main_BB1 : 0,
20     main_BB2 : 0,
21     main_BB3 : 0,
22     main_BB4 : 0,
23     main_BB5 : 0,
24     main_BB6 : 0,
25     main_BB7 : 0;
26
27 TRANSITION main_DROP_3
```

```
28    CONSUME
29      main_BB3 : 1;
30    PRODUCE
31      MUTEX_0 : 1,
32      main_BB4 : 1;
33  TRANSITION main_DROP_4
34    CONSUME
35      main_BB4 : 1;
36    PRODUCE
37      MUTEX_0 : 1,
38      main_BB5 : 1;
39  TRANSITION main_DROP_6
40    CONSUME
41      main_BB6 : 1;
42    PRODUCE
43      MUTEX_0 : 1,
44      main_BB7 : 1;
45  TRANSITION main_DROP_UNWIND_3
46    CONSUME
47      main_BB3 : 1;
48    PRODUCE
49      MUTEX_0 : 1,
50      main_BB6 : 1;
51  TRANSITION main_RETURN
52    CONSUME
53      main_BB5 : 1;
54    PRODUCE
55      PROGRAM_END : 1;
56  TRANSITION main_UNWIND_7
57    CONSUME
58      main_BB7 : 1;
59    PRODUCE
60      PROGRAM_PANIC : 1;
61  TRANSITION std_sync_Mutex_T_lock_0_CALL
62    CONSUME
63      MUTEX_0 : 1,
64      main_BB1 : 1;
65    PRODUCE
66      main_BB2 : 1;
67  TRANSITION std_sync_Mutex_T_lock_1_CALL
68    CONSUME
69      MUTEX_0 : 1,
70      main_BB2 : 1;
```

```
71    PRODUCE
72      main_BB3 : 1;
73  TRANSITION std_sync_Mutex_T_new_0_CALL
74    CONSUME
75      PROGRAM_START : 1;
76    PRODUCE
77      main_BB1 : 1;
```

Listing 5.1: The LoLA output for the program in Listing 4.4.

5.2.3. Test structure

The test programs are in the folder `examples/programs`. For each test program, there is a folder in `examples/results` that contains the three files `net.dot`, `net.pnml`, and `net.lola`.

The tests are grouped into categories:

- Basic: For basic programs like "Hello, World!" and a simple arithmetic calculator.
- Condvar: For programs concerning condition variables.
- Function call: For programs that test different types of function calls seen in Sec. 4.2.
- Mutex: For programs that use mutexes.
- Statement: For programs that test specific constructs such as a `match`, an infinite loop, an `Option`, a call to `panic!`, or `std::process::abort`.
- Thread: For programs involving multiple threads.

The structure of the folders in `examples/` mimics the file structure of the integration tests in `tests/`. As usual, the whole test suite can be run with the `cargo test` command.

5.2.4. Test implementation

The integration tests rely on the crates `assert_cmd`, `assert_fs`, and `predicates`. The idea to verify the output of the program by invoking the binary directly was taken from a specialized book for building CLI applications in Rust [Rust CLI WG, 2023, Chap. 1.6]. It was a useful resource for experimenting with `clap` to parse arguments as well.

Additionally, the integration tests use a shared submodule [Klabnik and Nichols, 2023, Chap. 11.3] that contains two convenient macros that save us from writing nearly all of the boilerplate code. These macros were defined using [Wirth and Keep, 2023] as a primary reference and with inspiration provided by [Oaten, 2023].

Listing 5.2 shows the macro responsible for generating the translation tests, while Listing 5.3 provides an example of how it is applied in the repository. For the sake of completeness, Listing 5.4 depicts the function used for the translation tests.

```
macro_rules! generate_tests_for_example_program {
    ($program_path:literal, $result_folder_path:literal) => {
        #[test]
        fn generates_correct_output_files() {
            super::utils::assert_output_files($program_path, $result_folder_path);
        }
    };
}
```

Listing 5.2: The macro that generates the translation tests.

```
mod utils;

mod calculator {
    super::utils::generate_tests_for_example_program!(
        "./examples/programs/basic/calculator.rs",
        "./examples/results/basic/calculator/"
    );
}

mod greet {
    super::utils::generate_tests_for_example_program!(
        "./examples/programs/basic/greet.rs",
        "./examples/results/basic/greet/"
    );
}

mod hello_world {
    super::utils::generate_tests_for_example_program!(
        "./examples/programs/basic/hello_world.rs",
        "./examples/results/basic/hello_world/"
    );
}
```

Listing 5.3: The contents of the file `basic.rs` listing all translation tests in the basic category.

```rust
pub fn assert_output_files(source_code_file: &str, output_folder: &str) {
    let mut cmd = Command::cargo_bin("cargo-check-deadlock").expect("Command not found");

    // Current workdir is always the project root folder
    cmd.arg("check-deadlock")
        .arg(source_code_file)
        .arg(format!("--output-folder={output_folder}"))
        .arg("--dot")
        .arg("--pnml")
        .arg("--filename=test")
        .arg("--skip-analysis");

    cmd.assert().success();

    for extension in ["lola", "dot", "pnml"] {
        let output_path = PathBuf::from(format!("{output_folder}test.{extension}"));
        let expected_output_path = PathBuf::from(format!("{output_folder}net.{extension}"));

        let file_contents =
            std::fs::read_to_string(&output_path).expect("Could not read output file to
            ↪  string");

        let expected_file_contents = std::fs::read_to_string(&expected_output_path)
            .expect("Could not read file with expected contents to string");

        if file_contents != expected_file_contents {
            panic!(
                "The contents of {} do not match the contents of {}",
                output_path.to_string_lossy(),
                expected_output_path.to_string_lossy()
            );
        }

        std::fs::remove_file(output_path).expect("Could not delete output file");
    }
}
```

Listing 5.4: The function that verifies the contents of the output files.

5.3. Visualizing the result

Visualizing the result is essential to understand the result of the deadlock detection. Thus, we invested time in researching different ways of achieving the same result, with and without a local installation required to make it as user-friendly as possible.

These instructions can also be found in the `README`[7] of the repository.

5.3.1. Locally

To see the MIR representation of the source code, the code can be compiled with the corresponding flag: `rustc --emit=mir <path_to_source_code>`

It is important to note that the nightly toolchain may produce different MIR compared to the stable version of the compiler. The reader is referred to Sec. 3.2.2 for more information.

To graph a net in `.dot` format, install the `dot` tool following the instructions on the GraphViz website[8].

Run `dot -Tpng net.dot -o outfile.png` to generate a PNG image from the resulting `.dot` file.

Run `dot -Tsvg net.dot -o outfile.svg` to generate a SVG image from the resulting `.dot` file.

More information and other possible image formats can be found in the documentation[9].

5.3.2. Online

To see the MIR representation of the source code, the Rust Playground[10] may be used.

The option "MIR" instead of "Run" must be selected in the dropdown menu. Note that the nightly version should be used instead of the stable version of *rustc*.

To graph a given DOT result, the Graphviz Online tool[11] offers a reliable alternative to the locally-installed tools. Alternatives exist such as Edotor[12] or SketchViz[13].

[7] https://github.com/hlisdero/cargo-check-deadlock/blob/main/README.md
[8] https://graphviz.org/download/
[9] https://graphviz.org/doc/info/command.html
[10] https://play.rust-lang.org/
[11] https://dreampuf.github.io/GraphvizOnline/
[12] https://edotor.net/
[13] https://sketchviz.com/new

5.3.3. Debugging

The program supports the verbosity flags defined in the crate `clap_verbosity_flag`[14]. For example, running the program with the flag `-vvv` prints debug messages that can be useful for pinpointing which line of the MIR is not being translated correctly.

The tool should then be invoked as follows:

```
cargo check-deadlock <path_to_program>/rust_program.rs -vvv
```

The model checker LoLA supports printing a "witness path" that shows a sequence of transition firings leading to a deadlock. This is extremely useful when extending the translator and the PN does not match the expected result for a given program.

A convenient script named `run_lola_and_print_witness_path.sh` is included in the repository to print the witness path for a `.lola` file. Fig. 5.1 illustrates the result of running the script on the file shown in Listing 4.5.

5.4. Integrating LoLA to the solution

As stated in Sec. 2.5.3, LoLA is the chosen model checker for this thesis. It acts as a backend that is responsible for verifying the absence of deadlocks. Integrating it was unfortunately not trivial.

5.4.1. Compilation

First, the compilation from the source code did not work on the hardware at our disposal. Changes to the code were necessary as newer versions of the C++ compiler tend to be more strict and reject or generate warnings for code that was previously accepted. Besides, one of the dependencies, kimwitu++[15], must be compiled from the source code too since it is not packaged for Linux distributions.

To preserve a working copy of the model checker for the future, indispensable for performing the deadlock analysis, a mirror[16] was created on GitHub where detailed instructions are provided for users. This aims to make the installation from source as straightforward as possible.

[14] https://docs.rs/clap-verbosity-flag/latest/clap_verbosity_flag/
[15] https://www.nongnu.org/kimwitu-pp/
[16] https://github.com/hlisdero/lola

```
lola found in $PATH.
lola: NET
lola:   reading net from examples/results/mutex/double_lock_deadlock/net.lola
lola:   finished parsing
lola:   closed net file examples/results/mutex/double_lock_deadlock/net.lola
lola:   20/65536 symbol table entries, 0 collisions
lola:   preprocessing...
lola:   finding significant places
lola:   11 places, 9 transitions, 9 significant places
lola:   computing forward-conflicting sets
lola:   computing back-conflicting sets
lola:   14 transition conflict sets
lola: TASK
lola:   read: EF ((DEADLOCK AND (PROGRAM_END = 0 AND PROGRAM_PANIC = 0)))
lola:   formula length: 59
lola:   checking reachability
lola:   Planning: workflow for reachability check: search (--findpath=off)
lola: STORE
lola:   using a bit-perfect encoder (--encoder=bit)
lola:   using 36 bytes per marking, with 0 unused bits
lola:   using a prefix tree store (--store=prefix)
lola: SEARCH
lola:   using reachability  graph (--search=depth)
lola:   using reachability preserving stubborn set method with insertion algorithm (--stubborn=tarjan)
lola: RUNNING
lola: RESULT
lola:   result: yes
lola:   The predicate is reachable.
lola:   3 markings, 2 edges
lola:   print witness path (--path)
lola:   writing witness path to stdout
std_sync_Mutex_T_new_0_CALL
std_sync_Mutex_T_lock_0_CALL
lola:   closed witness path file stdout
```

Figura 5.1: LoLA witness path output for the program in Listing 4.4.

5.4.2. Invoking the model checker

The second difficulty is that LoLA is compiled to an executable, not as a library, so our tool could not link to it. Instead, we are compelled to execute the binary from our `cargo-check-deadlock` binary to run the model checker passing the correct arguments[17]. The LoLA executable is part of the repository because it is needed to run the integration tests in the CI/CD pipeline using GitHub Actions. A user may also copy this executable to install LoLA in lieu of compiling it from scratch.

In the end, the user is responsible for installing the model checker separately to allow our tool to invoke it. A script `copy_lola_executable_to_cargo_home.sh` included in the repository

[17] https://github.com/hlisdero/cargo-check-deadlock/blob/main/src/model_checker/lola.rs

facilitates the task of copying the file to a folder that is already in the `$PATH`. We also considered other possibilities but none were feasible:

1. Using build scripts (`build.rs`) as described in the Cargo Book [Rust Project, 2023a, Chap. 3.8].
2. Modifying LoLA to turn it into a library.
3. Move a pre-compiled executable to the installation folder when running `cargo install`.
4. Define LoLA as a binary in the `Cargo.toml` [Rust Project, 2023a, Chap 3.2.1] and, hopefully, it gets moved to the cargo bin directory.
5. Define LoLA as an example in the `Cargo.toml` [Rust Project, 2023a, Chap 3.2.1] and, hopefully, it gets moved to the cargo bin directory.
6. Use a general-purpose build tool like `make`.

In the process of solving this second problem, we learned that cargo is mainly suited to dealing with dependencies expressed as Rust crates, which should be compiled when installed, not with arbitrary assets. In short, it is *not* meant to be a general-purpose build tool like `make`.

5.4.3. Expressing the property to check

The third challenge is finding a Computational Tree Logic* (CTL*) formula to instruct LoLA to search for deadlocks. Luckily, we can reuse the formula found in [Meyer, 2020]:

$$\text{EF (DEADLOCK AND (PROGRAM\_END} = 0 \text{ AND PROGRAM\_PANIC} = 0))$$

The formula represents the property to check for. It should be emphasized that not all deadlocks are interesting for our analysis. Our objective is to identify instances of deadlocks where the program execution is *unexpectedly* blocked. This scenario aligns with a dead PN as seen in Definition 14, where no transition is enabled, and the PN reaches a final state. However, we must exercise caution as there are cases where the PN is *expectedly* dead, such as when the program terminates or panics. These are states where execution normally halts. Thus, if we reach either the `PROGRAM_END` or the `PROGRAM_PANIC` place, the execution was successful, not a deadlock in the sense of Sec. 1.4.1. In conclusion, we exclude the `PROGRAM_END` and the `PROGRAM_PANIC` places by requiring them to be *unmarked* for the deadlock condition to hold. This is expressed by the "$= 0$" in the CTL* formula.

Lastly, we need to consider the temporal aspect. To specify that our state property eventually holds and to find a relevant path, we can utilize the "EF" operators in combination. The "F" stands for "eventually" and the "E" is the existential path quantifier [Meyer, 2020]. So the formula reads as:

"There exists eventually a path such that DEADLOCK (no transition may fire) and the place PROGRAM_PANIC has zero tokens and the place PROGRAM_END has zero tokens"

Other formulas may be constructed to check other properties. For this work, we take this formula as a given and we leave the user the possibility of checking other properties if she so desires. For a brief introduction to CTL*, see [Meyer, 2020].

5.5. Notable test programs

To wrap up this chapter, we introduce two noteworthy test programs that illustrate the current capabilities of the tool developed in this thesis. Our intention is to inspire others to contribute to this project or, at the very least, generate interest in the field of model checking.

First, Listing 5.5 showcases a simple version of the famous Dining Philosophers Problem proposed by Dijkstra. This version, affectionately nicknamed "Dating Philosophers", has only two philosophers and two forks on the table. A mutex needs to be locked to access each fork. When the philosophers try to grab both forks to eat, the program deadlocks, which is easy to verify by inspection. This deadlock is successfully detected by the tool. Moreover, a more complex version with 5 philosophers, for which the deadlock is also detected, is included in the repository[18]. It was omitted here due to the space constraints.

Second, observe the program in Listing 5.6. It models the classical producer-consumer problem. It uses a condition variable and a buffer with capacity for a single element. The access to the buffer is protected by a mutex. The producer generates 10 elements sequentially and the consumer processes them as they become available. The tool successfully verifies the absence of deadlock in the program.

[18] https://github.com/hlisdero/cargo-check-deadlock/blob/main/examples/programs/thread/dining_philosophers.rs

```rust
use std::sync::{Arc, Mutex};
use std::thread;

fn main() {
    let fork0 = Arc::new(Mutex::new(0));
    let fork1 = Arc::new(Mutex::new(1));

    let philosopher0 = {
        let left_fork = fork0.clone();
        let right_fork = fork1.clone();
        thread::spawn(move || {
            let _left = left_fork.lock().unwrap();
            let _right = right_fork.lock().unwrap();
        })
    };

    let philosopher1 = {
        let left_fork = fork1.clone();
        let right_fork = fork0.clone();
        thread::spawn(move || {
            let _left = left_fork.lock().unwrap();
            let _right = right_fork.lock().unwrap();
        })
    };

    // Wait for all threads to finish
    philosopher0.join().unwrap();
    philosopher1.join().unwrap();
}
```

Listing 5.5: A reduced version of the dining philosophers problem that deadlocks.

```rust
use std::sync::{Arc, Condvar, Mutex};
use std::thread;

fn main() {
    let buffer = Arc::new((Mutex::new(0), Condvar::new(), Condvar::new()));

    let producer_buffer = buffer.clone();
    let consumer_buffer = buffer.clone();

    let _producer = thread::spawn(move || {
        for i in 1..10 {
            let (lock, cvar_producer, cvar_consumer) = &*producer_buffer;
            let mut buffer = lock.lock().unwrap();

            while *buffer != 0 {
                buffer = cvar_producer.wait(buffer).unwrap();
            }

            *buffer = i;
            println!("Produced: {}", i);

            cvar_consumer.notify_one();
        }
    });

    let _consumer = thread::spawn(move || loop {
        let (lock, cvar_producer, cvar_consumer) = &*consumer_buffer;
        let mut buffer = lock.lock().unwrap();

        while *buffer == 0 {
            buffer = cvar_consumer.wait(buffer).unwrap();
        }

        let item = *buffer;
        *buffer = 0;
        println!("Consumed: {}", item);

        cvar_producer.notify_one();
    });
}
```

Listing 5.6: A solution to the producer-consumer problem.

Capítulo 6

Trabajos futuros

6.1. Reducing the size of the Petri net in postprocessing

[Murata, 1989] describes in a Section titled "Simple Reduction Rules for Analysis" six operations that preserve the properties of safeness, liveness, and boundedness of PN. See Definitions 13, 14 and 12 respectively for a refresher of what these properties mean.

The six operations involve simplifications that reduce the number of places or transitions in the Petri net. Next, we reproduce the names used for the reduction rules in the paper and Fig. 6.1 depicts the transformation that takes place in each case.

a) Fusion of Series Places.

b) Fusion of Series Transitions.

c) Fusion of Parallel Places.

d) Fusion of Parallel Transitions.

e) Elimination of Self-Loop Places.

f) Elimination of Self-Loop Transitions.

As these operations do not impact the liveness property, the outcome of the deadlock detection remains unchanged. Consequently, it might be advantageous to reduce the size of the PN after the translation process using specific methods available in the `netcrab` library. This step should be performed after translating all threads but before invoking the model checker.

Incorporating this functionality into the PN library itself would be more suitable, as it would allow other applications to benefit from this feature. It would be interesting to investigate whether this approach proves helpful when translating larger programs that contain hundreds or thousands of places and transitions.

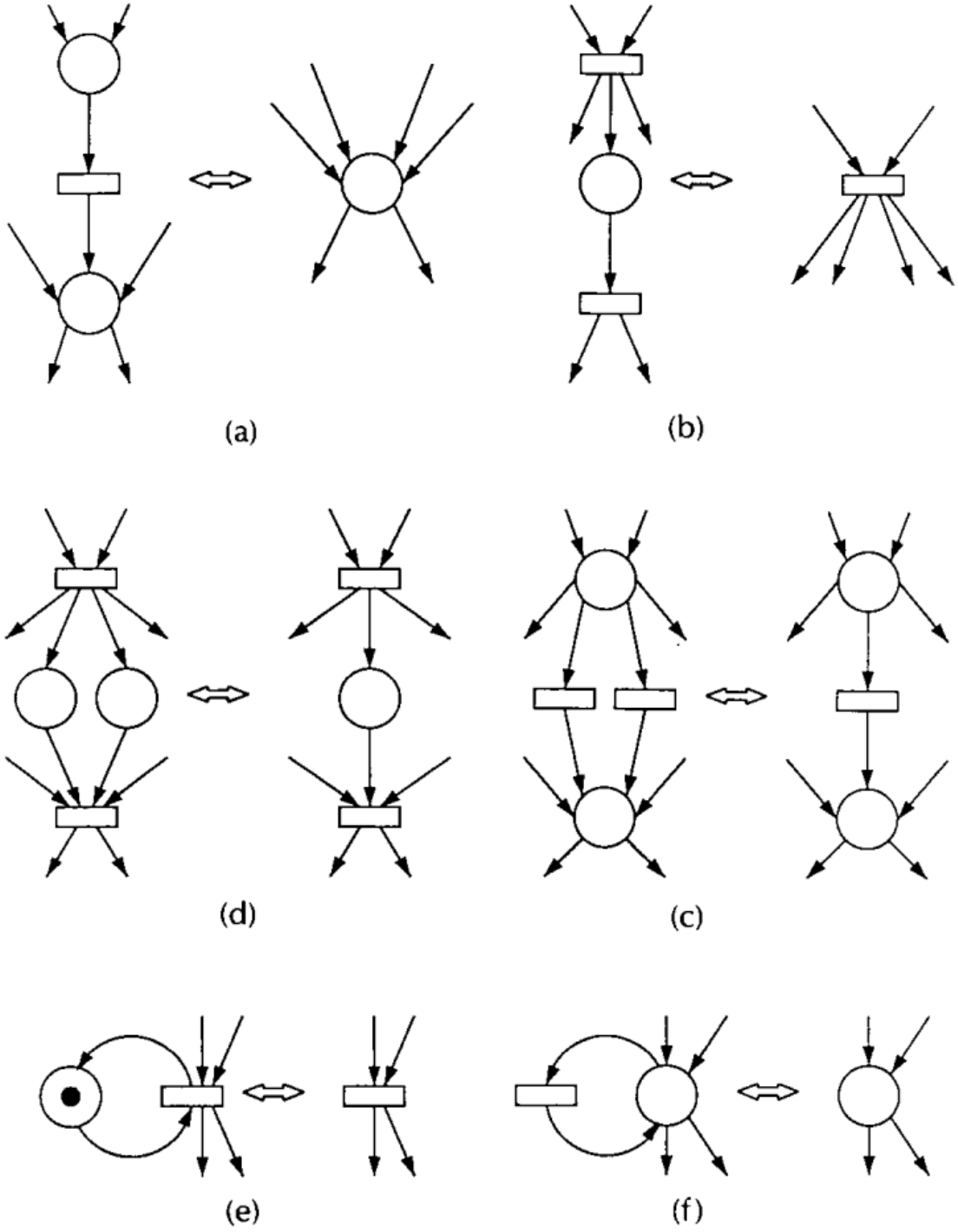

Figura 6.1: The reduction rules presented in Murata's paper.

One notable drawback of applying these operations is that it could obscure the source of the deadlock. It is valuable for the user to have precise information about the line in the source code

where the deadlock occurs. If the corresponding transitions or places representing this line are merged, this information is lost. However, this disadvantage may be deemed acceptable when dealing with extensive models, and the feature could be enabled or disabled at the discretion of the user.

6.2. Eliminating the cleanup paths from the translation

The error handling mechanism in the MIR must account for every possible scenario of failure during runtime. The aim of the *rustc* compiler is to ensure that compiled code fails gracefully, even in the most extreme circumstances, e.g., when the program is running out of memory or system calls fail unexpectedly due to hard limits on the resources available or other causes. However, the majority of this safeguarding cleanup code is never executed in practice. OOM errors and OS failures are uncommon and if they indeed emerge, a deadlock in user code is the least of our problems.

[Meyer, 2020] argues that the program will always terminate in a panic end-state once a single function call or assertion fails. Instead of translating the alternative path that the execution follows in the MIR, he proposes to set a token in the place `PROGRAM_PANIC` directly. This is equivalent to ignoring the specific cleanup target BB during the translation process and connecting the BB to the `PROGRAM_PANIC` place as if it were an `Unwind` terminator (Sec. 4.4.3).

This reduces the size of the Petri net model substantially. It comes with the disadvantage that cleanup BB are visited but never connected to other BB. These must be removed in a postprocessing step to not clutter the final model. Meyer's implementation does not seem to have performed this crucial step. It is unclear whether the implementation matches what the thesis proposed because the source code cannot be compiled anymore and no output examples are present in the repository[1].

The claim that the panic state is unrecoverable necessitates thorough examination, as we have previously observed in the introduction to Rust that the programmer has the option to utilize `std::panic::catch_unwind`. Furthermore, this intuitive reasoning might overlook situations in which a deadlock arises following a panic. This need not be a catastrophic failure. Consider for instance a thread that deadlocks while waiting on a message from another thread that panicked due to incorrect user input.

In conclusion, this modification of the translation logic looks promising to significantly reduce the number of places and transitions in the PN, especially in larger models, but more research is needed.

[1] https://github.com/Skasselbard/Granite

6.3. Translated function cache

A cache that stores functions after translating them is an interesting optimization to explore. The goal is to avoid redundant translations of the same function when it is called multiple times within the program. This idea was already briefly mentioned (but not implemented) in [Meyer, 2020]. The current implementation does not incorporate such caching mechanisms.

This cache would have to store a separate PN for each function. It could be realized as a `HashMap<rustc_hir::def_id::DefId, PetriNet>`, analogous to the function counter already present in the implementation. Furthermore, the translation process would need to merge/connect the Petri nets resulting from each translated function. This merging step requires support from the PN library `netcrab` to combine the multiple subnets into a cohesive whole.

However, connecting the individual Petri nets is not a trivial task, as a function may call an arbitrary number of other functions. Consequently, determining the appropriate "contact points" where the subnet should be connected becomes a challenging endeavor. The potential existence of numerous contact points, arising from the varying function call patterns, thus complicates the merging process.

Additionally, the Petri nets for each function should have labels that are unequivocal across the whole program or at least when exporting them to the format for the model checker. This requires generating slightly different versions of the same function for every call, which partly neglects the benefits of having a cache in the first place.

Lastly, some functions may not be cached at all in the case that special side effects exist. This happens for instance for all synchronization primitives currently supported. Their translation must be handled individually.

6.4. Recursion

Recursion in function calls poses a challenge in PN when defined as in Definition 1 due to the inability to properly map the data values to the model. PN lack the necessary expressive power to represent this compactly.

The number of times a recursive function is called ultimately depends on the data it is called with and cannot be determined at compile time. In normal program execution, a recursive function is pushed onto a new stack frame repeatedly until the base case is reached or the stack overflows. However, in PN, the function call where the base case is reached cannot be distinguished from the others, unless somehow the tokens representing recursion levels are distinct.

[Meyer, 2020, Sec. 3.4.2] discusses this problem and proposes using high-level Petri nets, i.e., Colored Petri nets (CPN) to solve it. High-level Petri nets provide a possible solution by allowing the distinction between tokens and the annotation of tokens with corresponding recursion levels.

Nevertheless, this necessitates a serious reconsideration of the entire translation logic owing to the different formalism. When using CPN each transition becomes a generalized function of input tokens of a specific type that generates tokens of the same or a different type. The resulting Petri net is substantially more complex and not all model checkers support CPN.

Mitigation strategies provide no comfort in this case either. On one hand, one could try to detect recursion and stop the translation, but recursion may exhibit unusual patterns that are not trivial to detect. For instance, consider a function A that calls a function B that calls a function C which finally calls A again. This recursion cycle may be arbitrarily long and adding this capability to the translator does not add much value compared to simply ignoring the problem and reaching a stack overflow.

On the other hand, [Meyer, 2020] suggests modeling each recursion level up to a maximum fixed depth, but this would impact verification results, as the properties of programs could vary with different maximum recursion depths. For every maximum recursion depth N, a counterexample program can be constructed that exhibits a different behavior, e.g., a deadlock, at recursion depth $N + 1$, hence avoiding detection.

6.5. Improvements to the memory model

Despite the seemingly straightforward implementation, devising a memory model that works in all cases is a challenging task. That being said, the current model is primarily a good first approximation and the solution has its drawbacks too.

Passing variables between MIR functions is not supported yet. This is a major drawback since it needs to be solved to support calling methods in `impl` blocks that receive synchronization variables. For this thesis, it was sufficient to write the programs in a simplified way to avoid this limitation but in a real case, this is not feasible.

There is significant coupling between the functions that handle the calls to functions in the `std::sync` module of the standard library and the `Memory`. A more generalized interface could be useful to add support for external libraries.

The idea of "linking" works well but does not match the semantics of Rust programs. In the long run, it would be preferable to delete the mapping if the variable gets moved to a different function. Taking references should also be treated as a distinct case from simply copying or using the variable.

The initial size of the `std::collections::HashMap` could be optimized for the average number of local variables in a typical MIR function. This could be a configuration parameter for the tool.

6.6. Higher-level models

The field of higher-level Petri net models is vast and encompasses numerous branches and potential methodologies. Exploring this domain offers a wide range of possibilities for advancing the modeling capabilities.

One notable advancement lies in the utilization of Colored Petri nets (CPN). Data values could then be modeled as tokens of different types, thereby enhancing the expressiveness and accuracy of the Petri net representation. A related paper in this regard is presented in the next chapter. [Meyer, 2020] also mentioned higher-level models when discussing improvements to his Petri net semantics for Rust. For an introduction to higher-level Petri nets, see [Murata, 1989]

Another intriguing addition to the current Petri net model involves the incorporation of inhibitor arcs. These arcs provide a means to model conditions in the source code where the presence of a zero value is checked. By introducing inhibitor arcs, Petri nets can effectively capture situations where the absence of a specific token is required for a transition to occur. For example, when checking a boolean flag used as a condition for a condition variable. Inhibitor arcs raise the expressive power of Petri nets to the level of Turing machines [Peterson, 1981].

Capítulo 7

Trabajos relacionados

In [Rawson and Rawson, 2022], the authors propose a generalized model based on colored Petri nets and implement an open-source middleware framework in Rust[1] to build, design, simulate and analyze the resulting Petri nets.

Colored Petri nets (CPN) are a type of Petri net that can represent more complex systems than traditional Petri nets. In a CPN, tokens have a specific value associated with them, which can represent various attributes or properties of the system being modeled. This allows for more detailed and accurate modeling of real-world systems, including those with complex data structures and behaviors. In the visual representation, each token has a color (analogous to a type in programming languages) and the transitions expect tokens from a particular color (type) and can generate tokens of the same color or tokens of a different color. As a short example, consider a transition with two input places and one output place representing the mixing of primary colors. If the input token colors are red and blue, then the output token color is purple. If the input token colors are yellow and blue, then the output token color is green.

The model proposed by the authors is an even more general type of Petri net, named Nondeterministic Transitioning Petri nets (NT-PN), which allows transitions to fire without having all their input places marked with tokens, while also allowing each transition to define which output places should be marked depending on the input. In other words, each transition defines arbitrary rules for its firing to take place. They explain briefly how the Petri net could be analyzed to solve for the maximal number of useful threads to execute the task modeled therein. They also mention the modeling step as a tool for checking for erroneous states before deploying an electronic or computer system.

In [De Boer et al., 2013], a translation from a formal language to Petri nets for deadlock detection in the context of active objects and futures is presented. The formal language chosen is Concurrent Reflective Object-oriented Language (Creol). It is an object-oriented modeling language designed for specifying distributed systems. In this paper, the program is made of

[1] https://github.com/MarshallRawson/nt-petri-net

asynchronously communicating active objects where futures are used to handle return values, which can be retrieved via a lock detaining `get` primitive (blocking) or a lock releasing `claim` primitive (non-blocking). After translating the program to a Petri net, reachability analysis is applied to detect deadlocks. This paper shows that a translation of asynchronous communication strategies to Petri nets with the goal of detecting deadlocks is also possible.

Capítulo 8

Conclusiones

This thesis has explored the translation of Rust programs into Petri net models for the purpose of deadlock and missed signal detection. Throughout the study, various aspects of the translation process have been examined, including the handling of function calls, threads, mutexes, and condition variables. The translator we developed has demonstrated its capability to accurately capture the concurrency and synchronization behavior of rather simple Rust programs.

The translation approach presented in this thesis has shown promising results, successfully modeling and detecting deadlocks in a range of test programs, comprising even two classical problems of concurrent programming. By harnessing the expressive power of Petri nets, the translator provides a visual representation of program behavior, facilitating the identification of potential synchronization issues. Most importantly, the translation produces a model that can be analyzed by a myriad of model checking tools, leveraging the existing academic work to bring solutions to industry problems. The incorporation of a succinct model for condition variables enhances the modeling capabilities and enables the detection of missed signals, which are a more intricate class of deadlock in concurrent systems.

Moving forward, there are several avenues for future research and improvement. One potential direction is the exploration of more complex programs and real-world applications to evaluate the scalability and effectiveness of the translation approach. Additionally, further refinement and optimization of the translation algorithms could enhance the efficiency of the analysis, specially higher-level models that would allow modeling the memory more effectively.

Overall, this thesis has made a significant contribution by developing a translator that bridges the gap between Rust programs and Petri nets. The insights gained from this research have shed light on the challenges and opportunities in modeling and analyzing concurrent systems at compile-time. Ideally, a programming language whose compiler detects concurrency problems would be a godsend for many applications. Building on the strengths of Petri nets, this possibility could be advanced further in the Rust programming language.

On a different note, the contribution of this thesis extends beyond the immediate benefits of

the proposed translator and its capabilities. By providing a solid, well-documented base for the translation of Rust programs into Petri nets, this work aims to make a meaningful contribution to the Rust community as a whole. It serves as a stepping stone for future endeavors, offering a reliable foundation upon which other tools and research projects can be built. It opens up new possibilities for exploring the analysis and verification of concurrent Rust programs using Petri nets. This, in turn, has the potential to drive further advancements in the field, stimulating innovation and promoting a deeper understanding of concurrent programming in Rust. With its comprehensive documentation and clear implementation, the translator not only facilitates immediate use but also serves as a valuable resource for those interested in studying or extending the translation techniques employed. Ultimately, this work aspires to ignite curiosity and inspire further contributions to the Rust ecosystem, fostering collaboration and growth in the community.

Bibliografía

[Aho et al., 2014] Aho, A. V., Lam, M. S., Sethi, R., and Ullman, J. D. (2014). *Compilers: Principles, Techniques, and Tools*. Pearson Education, 2 edition.

[Albini, 2019] Albini, P. (2019). RustFest Barcelona - Shipping a stable compiler every six weeks. https://www.youtube.com/watch?v=As1gXp5kX1M. Accessed on 2023-02-24.

[Arpaci-Dusseau and Arpaci-Dusseau, 2018] Arpaci-Dusseau, R. H. and Arpaci-Dusseau, A. C. (2018). *Operating Systems: Three Easy Pieces*. Arpaci-Dusseau Books, 1.00 edition. https://pages.cs.wisc.edu/~remzi/OSTEP/.

[Ben-Ari, 2006] Ben-Ari, M. (2006). *Principles of Concurrent and Distributed Programming*. Pearson Education, 2nd edition.

[Bernstein et al., 1987] Bernstein, P. A., Hadzilacos, V., Goodman, N., et al. (1987). *Concurrency control and recovery in database systems*, volume 370. Addison-Wesley Reading.

[Carreño and Muñoz, 2005] Carreño, V. and Muñoz, C. (2005). Safety verification of the small aircraft transportation system concept of operations. In *AIAA 5th ATIO and 16th Lighter-Than-Air Sys Tech. and Balloon Systems Conferences*, page 7423.

[Chifflier and Couprie, 2017] Chifflier, P. and Couprie, G. (2017). Writing parsers like it is 2017. In *2017 IEEE Security and Privacy Workshops (SPW)*, pages 80–92. IEEE.

[Coffman et al., 1971] Coffman, E. G., Elphick, M., and Shoshani, A. (1971). System deadlocks. *ACM Computing Surveys (CSUR)*, 3(2):67–78.

[Corbet, 2022] Corbet, J. (2022). The 6.1 kernel is out. https://lwn.net/Articles/917504/. Accessed on 2023-02-24.

[Coulouris et al., 2012] Coulouris, G., Dollimore, J., Kindberg, T., and Blair, G. (2012). *Distributed Systems, Concepts and Design*. Pearson Education, 5th edition.

[Czerwiński et al., 2020] Czerwiński, W., Lasota, S., Lazić, R., Leroux, J., and Mazowiecki, F. (2020). The reachability problem for petri nets is not elementary. *Journal of the ACM (JACM)*, 68(1):1–28. https://arxiv.org/abs/1809.07115.

[Davidoff, 2018] Davidoff, S. (2018). How Rust's standard library was vulnerable for years and nobody noticed. https://shnatsel.medium.com/how-rusts-standard-library-was-vulnerable-for-years-and-nobody-noticed-aebf0503c3d6. Accessed on 2023-02-20.

[De Boer et al., 2013] De Boer, F. S., Bravetti, M., Grabe, I., Lee, M., Steffen, M., and Zavattaro, G. (2013). A petri net based analysis of deadlocks for active objects and futures. In *Formal Aspects of Component Software: 9th International Symposium, FACS 2012, Mountain View, CA, USA, September 12-14, 2012. Revised Selected Papers 9*, pages 110–127. Springer.

[Dijkstra, 1964] Dijkstra, E. W. (1964). Een algorithme ter voorkoming van de dodelijke omarming. http://www.cs.utexas.edu/users/EWD/ewd01xx/EWD108.PDF.

[Dijkstra, 2002] Dijkstra, E. W. (2002). *Cooperating Sequential Processes*, pages 65–138. Springer New York, New York, NY.

[Esparza and Nielsen, 1994] Esparza, J. and Nielsen, M. (1994). Decidability issues for petri nets. *BRICS Report Series*, 1(8). https://tidsskrift.dk/brics/article/download/21662/19099/49254.

[Fernandez, 2019] Fernandez, S. (2019). A proactive approach to more secure code. https://msrc.microsoft.com/blog/2019/07/a-proactive-approach-to-more-secure-code/. Accessed on 2023-02-24.

[Gansner et al., 2015] Gansner, E. R., Koutsofios, E., and North, S. C. (2015). *Drawing Graphs With Dot*.

[Garcia, 2022] Garcia, E. (2022). Programming languages endorsed for server-side use at Meta. https://engineering.fb.com/2022/07/27/developer-tools/programming-languages-endorsed-for-server-side-use-at-meta/. Accessed on 2023-02-24.

[Gaynor, 2020] Gaynor, A. (2020). What science can tell us about C and C++'s security. https://alexgaynor.net/2020/may/27/science-on-memory-unsafety-and-security/. Accessed on 2023-02-24.

[Habermann, 1969] Habermann, A. N. (1969). Prevention of system deadlocks. *Communications of the ACM*, 12(7):373–ff.

[Hansen, 1972] Hansen, P. B. (1972). Structured multiprogramming. *Communications of the ACM*, 15(7):574–578.

[Hansen, 1973] Hansen, P. B. (1973). *Operating system principles*. Prentice-Hall, Inc.

[Heiner, 1992] Heiner, M. (1992). Petri net based software validation. *International Computer Science Institute ICSI TR-92-022, Berkeley, California*.

[Hillah and Petrucci, 2010] Hillah, L. M. and Petrucci, L. (2010). Standardisation des réseaux de Petri : état de l'art et enjeux futurs. *Génie logiciel : le magazine de l'ingénierie du logiciel et des systèmes*, 93:5–10.

[Hoare, 1974] Hoare, C. A. R. (1974). Monitors: An operating system structuring concept. *Communications of the ACM*, 17(10):549–557.

[Holt, 1972] Holt, R. C. (1972). Some deadlock properties of computer systems. *ACM Computing Surveys (CSUR)*, 4(3):179–196.

[Hosfelt, 2019] Hosfelt, D. (2019). Implications of Rewriting a Browser Component in Rust. https://hacks.mozilla.org/2019/02/rewriting-a-browser-component-in-rust/. Accessed on 2023-02-24.

[Howarth, 2020] Howarth, J. (2020). Why Discord is switching from Go to Rust. https://discord.com/blog/why-discord-is-switching-from-go-to-rust. Accessed on 2023-03-20.

[Huss, 2020] Huss, E. (2020). Disk space and LTO improvements. https://blog.rust-lang.org/inside-rust/2020/06/29/lto-improvements.html. Accessed on 2023-04-06.

[Jaeger and Levillain, 2014] Jaeger, E. and Levillain, O. (2014). Mind your language (s): A discussion about languages and security. In *2014 IEEE Security and Privacy Workshops*, pages 140–151. IEEE.

[Jannesari et al., 2009] Jannesari, A., Bao, K., Pankratius, V., and Tichy, W. F. (2009). Helgrind+: An efficient dynamic race detector. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pages 1–13. IEEE.

[Jüngel et al., 2000] Jüngel, M., Kindler, E., and Weber, M. (2000). The petri net markup language. *Petri Net Newsletter*, 59(24-29):103–104.

[Kani Project, 2023] Kani Project (2023). The Kani Rust Verifier. https://model-checking.github.io/kani/. Accessed on 2023-05-30.

[Karatkevich and Grobelna, 2014] Karatkevich, A. and Grobelna, I. (2014). Deadlock detection in petri nets: one trace for one deadlock? In *2014 7th International Conference on Human System Interactions (HSI)*, pages 227–231. IEEE.

[Kavi et al., 2002] Kavi, K. M., Moshtaghi, A., and Chen, D.-J. (2002). Modeling multithreaded applications using petri nets. *International Journal of Parallel Programming*, 30:353–371.

[Kavi et al., 1996] Kavi, K. M., Sheldon, F. T., and Reed, S. (1996). Specification and analysis of real-time systems using csp and petri nets. *International Journal of Software Engineering and Knowledge Engineering*, 6(02):229–248.

[Kehrer, 2019] Kehrer, P. (2019). Memory Unsafety in Apple's Operating Systems. https://langui.sh/2019/07/23/apple-memory-safety/. Accessed on 2023-02-24.

[Klabnik and Nichols, 2023] Klabnik, S. and Nichols, C. (2023). *The Rust programming language*. No Starch Press. https://doc.rust-lang.org/stable/book/.

[Klock, 2022] Klock, F. S. (2022). Contributing to Rust: Bootstrapping the Rust Compiler (rustc). https://www.youtube.com/watch?v=oG-JshUmkuA. Accessed on 2023-04-08.

[Knapp, 1987] Knapp, E. (1987). Deadlock detection in distributed databases. *ACM Computing Surveys (CSUR)*, 19(4):303–328.

[Kordon et al., 2022] Kordon, F., Bouvier, P., Garavel, H., Hulin-Hubard, F., Amat., N., Amparore, E., Berthomieu, B., Donatelli, D., Dal Zilio, S., Jensen, P., Jezequel, L., He, C., Li, S., Paviot-Adet, E., Srba, J., and Thierry-Mieg, Y. (2022). Complete Results for the 2022 Edition of the Model Checking Contest. http://mcc.lip6.fr/2022/results.php.

[Kordon et al., 2021] Kordon, F., Hillah, L. M., Hulin-Hubard, F., Jezequel, L., and Paviot-Adet, E. (2021). Study of the efficiency of model checking techniques using results of the mcc from 2015 to 2019. *International Journal on Software Tools for Technology Transfer*.

[Küngas, 2005] Küngas, P. (2005). Petri net reachability checking is polynomial with optimal abstraction hierarchies. In *Abstraction, Reformulation and Approximation: 6th International Symposium, SARA 2005, Airth Castle, Scotland, UK, July 26-29, 2005. Proceedings 6*, pages 149–164. Springer. PDF available from public profile on ResearchGate.

[Levick, 2022] Levick, R. (2022). Rust Before Main - Rust Linz. https://www.youtube.com/watch?v=q8irLfXwaFM. Accessed on 2023-04-30.

[Lipton, 1976] Lipton, R. J. (1976). The reachability problem requires exponential space. *Technical Report 63, Department of Computer Science, Yale University*. http://cpsc.yale.edu/sites/default/files/files/tr63.pdf.

[Matsakis, 2016] Matsakis, N. (2016). Introducing MIR. https://blog.rust-lang.org/2016/04/19/MIR.html. Accessed on 2023-04-14.

[Mayr, 1981] Mayr, E. W. (1981). An algorithm for the general petri net reachability problem. In *Proceedings of the Thirteenth Annual ACM Symposium on Theory of Computing*, STOC '81, page 238–246, New York, NY, USA. Association for Computing Machinery.

[Meyer, 2020] Meyer, T. (2020). A Petri Net semantics for Rust. Master's thesis, Universität Rostock | Fakultät für Informatik und Elektrotechnik. https://github.com/Skasselbard/Granite/blob/master/doc/MasterThesis/main.pdf.

[Miller, 2019] Miller, M. (2019). Trends, Challenges, and Strategic Shifts in the Software Vulnerability Mitigation Landscape. https://www.youtube.com/watch?v=PjbGojjnBZQ. Accessed on 2023-02-24.

[Monzon and Fernandez-Sanchez, 2009] Monzon, A. and Fernandez-Sanchez, J. L. (2009). Deadlock risk assessment in architectural models of real-time systems. In *2009 IEEE International Symposium on Industrial Embedded Systems*, pages 181–190. IEEE.

[Moshtaghi, 2001] Moshtaghi, A. (2001). Modeling Multithreaded Applications Using Petri Nets. Master's thesis, The University of Alabama in Huntsville.

[Mozilla Wiki, 2015] Mozilla Wiki (2015). Oxidation Project. https://wiki.mozilla.org/Oxidation. Accessed on 2023-03-20.

[Murata, 1989] Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580. http://www2.ing.unipi.it/~a009435/issw/extra/murata.pdf.

[Nelson, 2022] Nelson, J. (2022). RustConf 2022 - Bootstrapping: The once and future compiler. https://www.youtube.com/watch?v=oUIjG-y4zaA. Accessed on 2023-04-08.

[Nichols et al., 1996] Nichols, B., Buttlar, D., and Farrell, J. (1996). *Pthreads programming: A POSIX standard for better multiprocessing*. O'Reilly Media, Inc.

[Oaten, 2023] Oaten, T. (2023). Rust's Witchcraft. https://www.youtube.com/watch?v=MWRPYBoCEaY. Accessed on 2023-04-08.

[Perronnet et al., 2019] Perronnet, F., Buisson, J., Lombard, A., Abbas-Turki, A., Ahmane, M., and El Moudni, A. (2019). Deadlock prevention of self-driving vehicles in a network of intersections. *IEEE Transactions on Intelligent Transportation Systems*, 20(11):4219–4233.

[Peterson, 1981] Peterson, J. L. (1981). *Petri Net Theory and the Modeling of Systems*. Prentice-Hall.

[Petri, 1962] Petri, C. A. (1962). Kommunikation mit Automaten. *Institut für Instrumentelle Mathematik*, 3. http://edoc.sub.uni-hamburg.de/informatik/volltexte/2011/160/.

[Rawson and Rawson, 2022] Rawson, M. and Rawson, M. (2022). Petri nets for concurrent programming. *arXiv preprint arXiv:2208.02900*.

[Reid, 2021] Reid, A. (2021). Automatic Rust verification tools (2021). https://alastairreid.github.io/automatic-rust-verification-tools-2021/. Accessed on 2023-02-20.

[Reid et al., 2020] Reid, A., Church, L., Flur, S., de Haas, S., Johnson, M., and Laurie, B. (2020). Towards making formal methods normal: meeting developers where they are. Accepted at HATRA 2020.

[Reisig, 2013] Reisig, W. (2013). *Understanding Petri Nets: Modeling Techniques, Analysis Methods, Case Studies*. Springer-Verlag Berlin Heidelberg, 1st edition.

[Rust CLI WG, 2023] Rust CLI WG (2023). Command Line Applications in Rust. https://rust-cli.github.io/book/. Accessed on 2023-06-08.

[Rust on Embedded Devices Working Group, 2023] Rust on Embedded Devices Working Group (2023). The Embedded Rust Book. https://docs.rust-embedded.org/book/. Accessed on 2023-06-02.

[Rust Project, 2023a] Rust Project (2023a). The Cargo Book. https://doc.rust-lang.org/cargo/. Accessed on 2023-06-08.

[Rust Project, 2023b] Rust Project (2023b). The rustc Book. https://doc.rust-lang.org/rustc/. Accessed on 2023-02-20.

[Rust Project, 2023c] Rust Project (2023c). The Rustonomicon. https://doc.rust-lang.org/nomicon/. Accessed on 2023-04-19.

[Rust Project, 2023d] Rust Project (2023d). The rustup Book. https://rust-lang.github.io/rustup/index.html. Accessed on 2023-05-02.

[Rust Project, 2023e] Rust Project (2023e). The Unstable Book. https://doc.rust-lang.org/unstable-book/the-unstable-book.html. Accessed on 2023-04-14.

[Savage et al., 1997] Savage, S., Burrows, M., Nelson, G., Sobalvarro, P., and Anderson, T. (1997). Eraser: A dynamic data race detector for multithreaded programs. *ACM Transactions on Computer Systems (TOCS)*, 15(4):391–411.

[Schmidt, 2000] Schmidt, K. (2000). Lola a low level analyser. In *Application and Theory of Petri Nets 2000: 21st International Conference, ICATPN 2000 Aarhus, Denmark, June 26–30, 2000 Proceedings 21*, pages 465–474. Springer.

[Shibu, 2016] Shibu, K. V. (2016). *Introduction to Embedded Systems*. McGraw Hill Education (India), 2nd edition.

[Silva and Dos Santos, 2004] Silva, J. R. and Dos Santos, E. A. (2004). Applying petri nets to requirements validation. *IFAC Proceedings Volumes*, 37(4):659–666.

[Simone, 2022] Simone, S. D. (2022). Linux 6.1 Officially Adds Support for Rust in the Kernel. https://www.infoq.com/news/2022/12/linux-6-1-rust/. Accessed on 2023-02-24.

[Singhal, 1989] Singhal, M. (1989). Deadlock detection in distributed systems. *Computer*, 22(11):37–48.

[Stack Overflow, 2022] Stack Overflow (2022). 2022 Developer Survey. https://survey.stackoverflow.co/2022/#section-most-loved-dreaded-and-wanted-programming-scripting-and-markup-languages. Accessed on 2023-02-22.

[Stepanov, 2020] Stepanov, E. (2020). Detecting Memory Corruption Bugs With HWASan. https://android-developers.googleblog.com/2020/02/detecting-memory-corruption-bugs-with-hwasan.html. Accessed on 2023-02-24.

[Stoep and Hines, 2021] Stoep, J. V. and Hines, S. (2021). Rust in the Android platform. https://security.googleblog.com/2021/04/rust-in-android-platform.html. Accessed on 2023-02-22.

[Stoep and Zhang, 2020] Stoep, J. V. and Zhang, C. (2020). Queue the Hardening Enhancements. https://android-developers.googleblog.com/2020/02/detecting-memory-corruption-bugs-with-hwasan.html. Accessed on 2023-02-24.

[Szekeres et al., 2013] Szekeres, L., Payer, M., Wei, T., and Song, D. (2013). Sok: Eternal war in memory. In *2013 IEEE Symposium on Security and Privacy*, pages 48–62. IEEE.

[The Chromium Projects, 2015] The Chromium Projects (2015). Memory safety. https://www.chromium.org/Home/chromium-security/memory-safety/. Accessed on 2023-02-24.

[The Rust Project Developers, 2019] The Rust Project Developers (2019). Rust case study: Community makes rust an easy choice for npm. https://www.rust-lang.org/static/pdfs/Rust-npm-Whitepaper.pdf.

[Thierry Mieg, 2015] Thierry Mieg, Y. (2015). Symbolic Model-Checking using ITS-tools. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 9035 of *Lecture Notes in Computer Science*, pages 231–237, London, United Kingdom. Springer Berlin Heidelberg.

[Thompson, 2023] Thompson, C. (2023). How Rust went from a side project to the world's most-loved programming language. https://www.technologyreview.com/2023/02/14/1067869/rust-worlds-fastest-growing-programming-language/.

[Toman et al., 2015] Toman, J., Pernsteiner, S., and Torlak, E. (2015). Crust: a bounded verifier for rust (n). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 75–80. IEEE.

[Van der Aalst, 1994] Van der Aalst, W. (1994). Putting high-level petri nets to work in industry. *Computers in industry*, 25(1):45–54.

[van Steen and Tanenbaum, 2017] van Steen, M. and Tanenbaum, A. S. (2017). *Distributed Systems*. Pearson Education, 3rd edition.

[Weber and Kindler, 2003] Weber, M. and Kindler, E. (2003). The Petri Net Markup Language. *Petri Net Technology for Communication-Based Systems: Advances in Petri Nets*, pages 124–144.

[Wirth and Keep, 2023] Wirth, L. and Keep, D. (2023). The Little Book of Rust Macros. https://veykril.github.io/tlborm/introduction.html. Accessed on 2023-06-08.

[Wu and Hauck, 2022] Wu, Y. and Hauck, A. (2022). How we built Pingora, the proxy that connects Cloudflare to the Internet. https://blog.cloudflare.com/how-we-built-pingora-the-proxy-that-connects-cloudflare-to-the-internet/. Accessed on 2023-03-20.

[Zhang and Liua, 2022] Zhang, K. and Liua, G. (2022). Automatically transform rust source to petri nets for checking deadlocks. *arXiv preprint arXiv:2212.02754*.