

Дипломная работа

Максим Борисяк

14 июня 2013 г.

1 Введение *** Need review ***

В последние годы во многих областях науки, таких как, например, бионформатика или экспериментальная физика, требуется обрабатывать все большее и большее количество данных. В качестве примера можно привести Большой Адронный Коллайдер (БАК). В результате одного эксперимента поток 'сырых' данных с детектора *ALICE* может превышать 1 гигабайт в секунду [???].

Для обработки больших объемов данных используются распределенные вычислительные системы, мощности которых могут достигать петафлопов. Так для обработки данных БАК используется сложная распределенная система *LHC Computing Grid*. Такие мощности достигаются за счет громадного количества вычислительных узлов. Например, суперкомпьютер Sequoia мощностью более 1 петафлопа имеет более миллиона ядер [???].

Однако с ростом объемов данных, и, как следствие, с ростом необходимых вычислительных ресурсов растет сложность задач эффективной обработки. Известно, что принципы, лежащие в основе эффективных алгоритмов, могут существенно усложняться при переходе к параллельным вычислениям и даже при изменении масштабов вычислительных систем.

Другой проблемой становится сложность создания самих распределенных алгоритмов, так как многие из них теряют наглядность из-за реализации большого количества технических аспектов, которые необходимо учитывать при их разработке. Более того, некоторые распределенные системы не имеют четких границ, так как задействованные в ней участники могут постоянно меняться, что дополнительно усложняет контроль за вычислениями. Так для ускорения обработки результатов БАК был задействован проект *LHC@home* [???], помочь которому мог любой желающий путем предоставления части вычислительной мощности собственного компьютера.

Одним из решений описанных выше проблем стала концепция *потоков данных (dataflow)*. Под потоком данных обычно понимают модель представления алгоритмов. Описание вычислительного процесса в такой концепции, обычно, представляет из себя следующие элементы:

- набор атомарных исполнителей-функций, реализованный вне концепции потоков данных, например, в виде программ или динамических библиотек;

- описание зависимостей между исполнителями — результат работы одних исполнителей передается на вход (в качестве аргументов) другим;
- описание стратегии поведения потока данных;
- иерархичность — иногда целый поток данных может быть представлен в виде одного исполнителя внутри другого потока данных.

Иными словами, поток данных представляется в виде наборов исполнителей, каждый из которых выполняет некое преобразование поступивших ему на вход данных, соединенных каналам передачи данных.

Концепция потоков данных обладает важным преимуществом — она полностью изолирует создателя алгоритма от необходимости описывать многие технические аспекты вычислений. В этом они родственны декларативным языкам программирования, в которых программы выглядят как наборы определений или описаний зависимостей. В частности, некоторые потоки данных можно без потери структуры преобразовать к программе, например, на языке Haskell [???].

Другим преимуществом описания вычислительных процессов в виде потоков данных являются широкие возможности для автоматического распределения ресурсов при запуске потока данных в распределенных средах, а также для предварительного анализа необходимых ресурсов для эффективного запуска. Эти свойства являются крайне востребованными для научных вычислений, так как объемы вычислений явно требуют больших ресурсов недостижимых на системах с одним потоком выполнения. Как в примере с проектом *LHC@home* потоки данных также хорошо описывают вычисления в динамических распределенных средах. Возможность предварительной оценки требуемых ресурсов важна для эффективной утилизации мощностей суперкомпьютеров, вычислительные способности которых могут заметно превосходить требуемые ресурсы. Так задаче получения таких оценок посвящена данная работа.

Родственной к потокам данных является концепция потоков управления (controlflow), основное отличие которых от потоков данных в том, что связи между исполнителями означают передачу управления, и часто исполнители имеют доступ к общему хранилищу данных. Программы в терминах потоков управления имеют явно указанные участки параллельного выполнения, создания и слияния параллельных ветвей.

Эти две концепции объединяют под названием поток работ (workflow). Среди достоинств концепции потока работ следует выделить следующие:

- изолированность программы, сценария или схемы от технических особенностей среды выполнения;
- наглядность программ — чаще всего поток работ можно представить в графической нотации близкой к блок-схемам;
- простота создания программ — создание программ, особенно в графической нотации, не требует специальных знаний из области информатики, что позволяет, например, физикам-экспериментаторам не тратить большое количество времени на изучение новых языков или сред обработки результатов.

В отличие от потоков управления потоки данных, как уже было упомянуто, ориентированы на научные вычисления. Среди особенностей можно выделить следующие:

- работа в распределенных средах сложной динамичной структуры, ресурсы которой могут быть заранее неизвестны;
- высокая отказоустойчивость,
- работа с большим объемом данных.

Заметим, что среды построения и запуска потоков данных значительно отличаются по внутренним механизмам от аналогичных сред для потоков управления из-за перечисленных выше отличий в концепциях.

Также стоит отметить, что потоки данных имеют много общего с *сетями процессов Кана (Kahn Process Networks)*, сходства и различия с которыми будет рассмотрено ниже.

Отметим также растущую популярность концепции потоков данных в научных кругах — так в описанном выше примере с БАК, для обработки результатов используются потоки данных или близкие к ним концепции [???].

1.1 Мотивация *** Need review ***

В данной работе рассматривается модель потоков данных, которая, по мнению автора, предоставляет наиболее естественное описание программ для научных расчетов в концепции потоков данных. Также рассматриваемая модель близка к моделям существующих и активно используемых сред для построения и запуска потоков данных.

В данной работе следует выделить несколько задач:

- построение формальной модели потока данных;
- анализ потока данных на наличие ошибок;
- оценка ресурсов и стратегий, требуемых для эффективного запуска потока данных в распределенных средах.

Главные алгоритмы были реализованы в виде модуля языка Python для возможности использования в интерактивном интерпретаторе для языка Python.

1.2 Обзор существующих моделей представления потоков данных *** Need review ***

Среди моделей представления потоков данных следует выделить следующие:

- программы на высокоуровневых языках программирования;
- графы;
- сети Петри;

- сети процессов Кана.

Как уже было сказано, концепция сетей процессов Кана имеет некие отличия от рассматриваемой в этой работе концепции потоков данных, однако определение потоков данных может варьироваться и обозначает скорее направление в представлении алгоритмов обработки, поэтому можно считать сети процессов Кана, если не одной из моделей представления потоков данных, то как минимум родственной моделью, и следовательно, заслуживающей рассмотрения как минимум для демонстрации возможных подходов.

1.2.1 Программы на высокоуровневых языках программирования

Любой вычислительный процесс можно наглядно описать в терминах современных языков программирования при должном уровне абстракции от технических особенностей среды выполнения. Наиболее подходящими языками для описания потоков данных являются скриптовые языки, например, Python, либо декларативные языки, например, Haskell или Erlang. Декларативная парадигма, как уже было сказано, достаточно близка концепции потоков данных, однако порой декларативные языки слишком строги для описания потоков данных.

Основной проблемой такого подхода является недостаточная наглядность для людей не знакомых с конкретным языком программирования и требует знания возможных тонкостей этого языка. Более того, обычно визуальное представление является более наглядным для человека, чем текстовальное. Этим можно объяснить меньшую популярность такого подхода.

В качестве примера системы, использующей языки программирования как представление потока данных, можно привести систему *GridAnd* [???].

1.2.2 Графы

Графы являются самой популярной моделью потоков данных. Главным достоинством является возможность интуитивно понятного визуального представления графов, которое не требует специальных знаний в области информатики.

Другим достоинством графов как модели представления потока данных является возможность построения иерархичных потоков данных, где один поток данных, представляется в виде вершины в потоке более высокого уровня, тем самым реализуя принципы модульности и абстракции.

Особо стоит выделить хорошо изученные свойства графов и множество алгоритмов для их анализа, а так же выразительную силу графов. Например, в данной работе с помощью графов описываются сами потоки данных, качественно поведение каждого блока (в виде конечного автомата), поведение блока во время запуска потока данных (также в виде автомата).

1.2.3 Сети Петри

Другой популярной концепцией представления потоков данных являются сети Петри, которые можно считать развитием идеи графов. Как и в случае с графами, сети Петри можно наглядно представить используя визуальные средства.

Также доступен широкий набор средств для анализа сетей Петри. Кроме того, в сетях Петри явно присутствуют элементы паралелизма, хотя сети Петри лучше подходят для описания потоков управления (чем возможно объясняется их популярность в системах построения потоков управления), нежели потоков данных, так как переход так называемых маркеров от одного узла к другому означает скорее передачу управления, нежели передачу данных. Однако представление потоков данных в виде сетей Петри также возможно.

1.2.4 Сети процессов Кана

Сети процессов Кана используются для описания потоков данных ориентированных на потоковую обработку, в отличие от рассматриваемой в этой работе модели, где данные подаются на вход поэлементно или одной 'порцией'. Формально в этой концепции каждый процесс (исполнитель) представляется функцией, которая преобразует входящие потоки данных целиком, и лишь в частном случае это эквивалентно поэлементной обработке. В таких случаях мы получим концепцию близкую к концепции потоков данных. Однако в сетях процессов Кана для разрешения возможных неопределенностей в поведении процессов накладываются ограничения на функции, которыми представляется процесс, вводятся правила поглощения и испускания данных; в концепции потоков данных вместо наложения ограничений на функции стараются анализировать саму схему потока данных, которая, так как не является потоковой, сама по себе может не допускать потенциальных неопределенностей [???].

Стоит также отметить, что сети процессов Кана, аналогично потокам данных, могут быть естественно представлены с помощью декларативных языков программирования. Например, используя функции преобразующие бесконечные списки в языке Haskell, можно добиться точного соответствия формального определению сетей процессов Кана.

1.3 Неформальное описание модели *** Need review ***

В качестве основы для построения модели были взяты основные принципы работы популярных сред построения и запуска потоков данных. Сам поток данных представляется в виде ориентированного графа, вершинами которого являются блоки. Блоки реализуют некий алгоритм преобразования данных и в зависимости от способа создания этого алгоритма делятся на два типа:

- атомарные — встроенные в систему построения и запуска потоков данных в виде внутренних подпрограмм;
- составные — блок, который сам представляется в виде потока данных и является вложенным потоком данных.

У каждого блока (кроме специальных или 'фиктивных') существуют наборы входных и выходных портов. Входные порты в некотором смысле являются 'аргументами' блока — по ним передаются данные для преобразования. Выходные порты имеют смысл результата выполнения преобразования — по ним данные передаются для дальнейшей обработки. Каждый

порт обычно обладает определенным смыслом, который определяется алгоритмом преобразования.

Ребро графа потока данных соединяет выходной порт одного блока с входным портом другого блока, задавая тем самым, последовательность обработки данных. Ребра можно также рассматривать как FIFO-каналы, передача по которым ради простоты будем считать мгновенной.

Естественно, что работа потока данных зависит от внутренних состояний программ его блоков. В данной работе алгоритмы блоков заменяются на конечные автоматы Мили, описывающие зависимости наборов выходных данных от входных. Таким образом мы получаем качественное описание работы блоков, что существенно упрощает задачу анализа, освобождая от необходимости описывать анализ алгоритмов блоков, который выходит за рамки рассматриваемых вопросов.

Стоит отметить, что обычно данные имеют тип, однако рассмотрение типов данных приводит к усложнению модели, не приводя при этом ни к каким дополнительным результатам.

1.4 Базовые понятия и определения *** Need review ***

В данном разделе будут введены базовые понятия, используемые в данной работе, в том числе, понятия атомарного блока, составного блока, потока данных, автомата Мили соответствующего блоку и так далее.

Основной единицей потока данных является атомарный блок.

Определение 1. *Шаблон атомарного блока s назовем кортеж имеющий два не пересекающихся набора портов: входной и выходной. $s = (FA, I, O)$, где:*

- $I = \{b_i^I | i = \overline{1, N_I}, N_I \in \mathbb{N}\}$ — конечное множество уникальных входных портов;
- $O = \{b_i^O | i = \overline{1, N_O}, N_O \in \mathbb{N}\}$ — конечное множество уникальных выходных портов;
- FA — конечный автомат Мили, смысл и структура которого будет обсуждаться в дальнейшем.

Сразу заметим, что не любая тройка (FA, I, O) является шаблоном атомарного блока, так как FA должен соответствовать множествам I и O .

Стоит отметить, что природа множеств O и I совершенно не важна, например, они могут быть просто множеством чисел: $I, O = \{i \in \overline{1, N_{I,O}}\}$, на практике часто используют строки для именования портов, так как порты могут существенно различаться по смыслу, а их именование помогает избежать путаницы. Последний подход используется в рисунках.

Определение 2. *Атомарным блоком $b = (id, s, state)$, где:*

- id — уникальный для всех блоков идентификатор;
- s — шаблон атомарного блока;
- $state \in S$, где S множество состояний FA_s — текущее состояние атомарного блока.

Идентификатор id введен из-за следующих соображений: в потоке данных некие блоки могут описываться одним и тем же шаблоном, находиться в одном и том же состоянии, но все равно будут являться различными блоками, в частности, могут перейти в различные состояния в следующие моменты времени либо могут иметь в потоке данных разный смысл и, соответственно, иметь различные соединения с другими блоками.

В реальных системах для построения и запуска потока данных шаблон атомарного блока представляется в виде подпрограммы или модуля (функции, класса некоего объектно-ориентированного языка программирования, динамической библиотеки), который используя данные некоего набора своих входных портов, вычисляет и записывает данные в свои выходные порты. В таком случае, атомарный блок представляется в виде набора внутренних переменных, либо (что тоже самое) в виде экземпляра класса объектно-ориентированного языка программирования соответствующего шаблону. В такой интерпретации некую вычислимую функцию $f(x_1, x_2, \dots, x_n)$ можно представить в виде шаблона блока b_f с входными портами x_1, x_2, \dots, x_n и одним выходным портом F , который при получении данных со всех входных портов выписывает значения функции $f(x_1, x_2, \dots, x_n)$ в порт F .

На рисунке 1 схематично изображен шаблон блока $Loop$, для которого $inputs(Loop) = \{xs, f\}$, $outputs(Loop) = \{fs, x\}$.

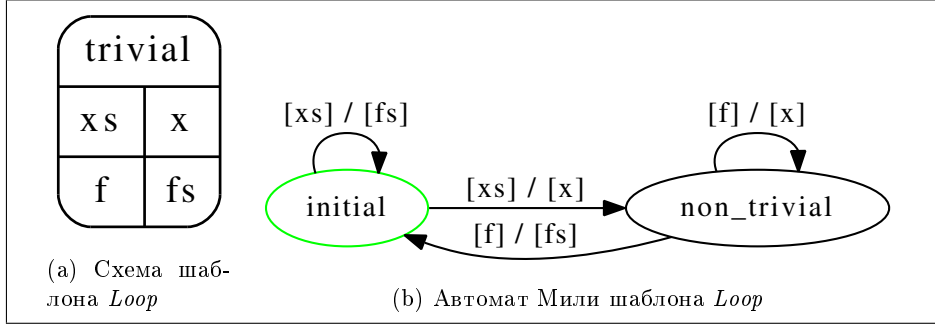


Рис. 1: Схематичное изображение шаблона блока $Loop$ (a) и соответствующего ему конечного автомата Мили (b).

Основной структурой для построения схемы потока данных является составной блок.

Определение 3. *Шаблоном составного блока будем называть кортеж $s = (B, E, I, O)$, где:*

- B - конечное множество блоков (сразу заметим, что блоком может являться и составной блок, определенный ниже).
- E - множество ребер вида $e_b = (b_1, b_1^O, b_2, b_2^I)$, $b_1, b_2 \in B$, $b_1^O \in outputs(b_1)$, $b_2^I \in inputs(b_2)$, либо вида $e_I = (c^I, b, b^I)$, $e_O = (b, b^O, c^O)$, где $b \in B$, $b^I \in inputs(b)$, $b^O \in outputs(b)$, $c^O \in O$, $c^I \in I$. Множества ребер e_b будем обозначать как E^B , e_I и e_O как E^I и E^O соответственно.
- I, O - наборы входных и выходных портов (по аналогии с атомарным блоком).

В дальнейшем в выражениях вида I_c или O_c нижний индекс будет указывать на схему блока, которому принадлежат эти множества.

Для анализа внутренней структуры удобно рассматривать иное представление шаблона составного блока c , добавляя фиктивные блоки STOCK, $I_{\text{STOCK}} = O_{\text{STOCK}} = O_c$, и SOURCE, $I_{\text{SOURCE}} = O_{\text{SOURCE}} = I_c$:

$$\hat{E}^I = \{(\text{SOURCE}, \text{SOURCE}^O, b, b^I) | (\text{SOURCE}^O, b, b^I) \in E^I\}$$

$$\hat{E}^O = \{(b, b^O, \text{STOCK}, \text{STOCK}^I) | (b, b^O, \text{STOCK}) \in E^O\}$$

$$\hat{c} = (B \cup \{\text{STOCK}, \text{SOURCE}\}, E^B \cup \hat{E}^I \cup \hat{E}^O)$$

В этом случае шаблон составного блока \hat{c} описывается графом с ребрами вида (u, u^O, v, v^I) , а входными и выходными портами всего шаблона считаются входные и выходные порты блоков SOURCE и STOCK. На рисунке 2 изображены примеры шаблонов составных блоков во второй интерпретации.

Определение 4. *Составным блоком* будем называть $b = (id, c = (B, E, I, O), state)$, где

- id — уникальный идентификатор;
- c — шаблон составного блока;
- $state \in \Sigma \times 2^E = \prod S(B) \times 2^E \times 2^B$ — текущее состояние составного блока, включающее состояние каждого блока (подпространство $\Sigma = \prod S(B)$), активную волну ($\omega \in \Omega = 2^E$) и множество активных блоков ($\vartheta \in \Theta = 2^B$), смысл которых будет пояснен позже.

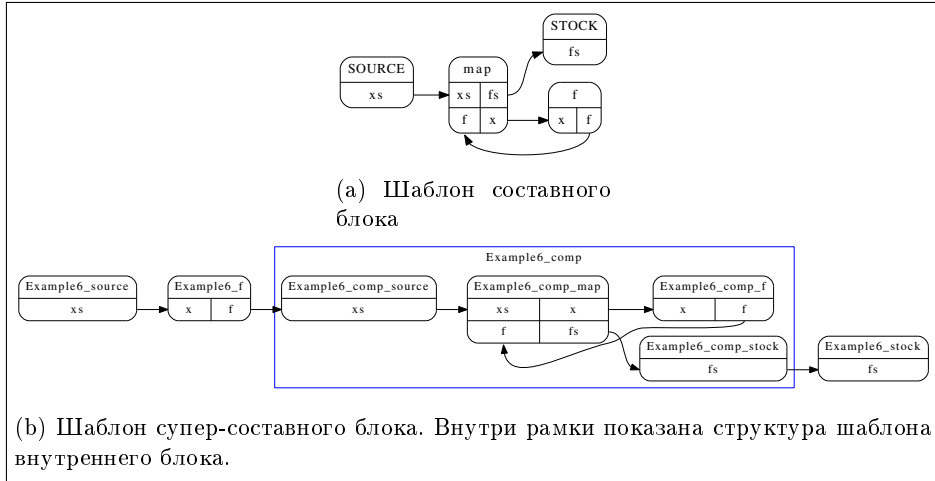


Рис. 2: Примеры графов составного и супер-составного блока.

Заметим, что составной блок (в исходной интерпретации) внешне не отличается от атомарного блока. Поэтому под понятием блока мы будем подразумевать либо атомарный блок, либо составной блок, различая их только в том случае, когда речь заходит о их внутреннем строении. Составной блок может включать в себя другие составные блоки. На рисунке 2b показан пример составного блока, включающего другой составной блок, иными

словами супер-составного блока. На практике свойство вложенности реализует принцип модульности программ — модулем в нашем случае будет шаблон составного блока. Бессмысленно рассматривать случаи, в которых структура составного блока бесконечна в глубину, например, в случае когда шаблон составного блока содержит блок того же шаблона. Поэтому в дальнейшем, мы будем рассматривать только составные блоки конечной глубины.

Заметим, что составным блоком можно описать любую программу, имея в распоряжении необходимые примитивы — шаблоны атомарных блоков, реализующие Тьюринг-полную систему операций. Доказательства этого факта выходят за задачи данной работы и может быть найдено, например, в работе [???]. Из-за достаточной выразительности составного блока, схему потока данных мы определим просто как схему некоего составного блока.

Теперь неформально поясним схему работы составного блока и смысл его структуры. В общепринятых определениях потока данных ребра графа шаблона составного блока обозначают зависимость по данным, то есть ребро $e = (u, u^O, v, v^I)$ представляет собой *FIFO*-канал соединения между портами блоков и можно словесно интерпретировать следующим образом: блок v в качестве данных из входного порта v^I должен использовать данные выходного порта u^O блока u при их наличии. Блок u после завершения своей работы может "испустить" данные по порту u^O которые мгновенно "переносятся" на все ребра из u^O блока u , если эти ребра свободны. Когда блоку v требуются данные из порта v^I , данные с ребра e мгновенно "переносятся" в соответствующий порт блока. В этом и есть смысл подпространства $\Omega = 2^E$ в определении состояния составного блока — *активной волной* ω мы назовем множество ребер, которые содержат на данный момент данные.

Однако такая схема работы составного блока порождает множество неопределенностей в работе некоторых составных блоков, например, неопределенность выбора ребра для "поглощения" данных из определенного порта (состояние гонки). Неопределенности подобного рода мы будем рассматривать как ошибки, обнаружению которых посвящена часть данной работы. Такого рода неоднозначности, особенно в реальных системах, то есть в условиях неопределенного времени вычислений блоков и их параллельного выполнения, может привести к совершенно иному ходу потока данных, нежели было задумано автором потока данных. Поэтому наличие подобного рода неопределенностей (состояний гонки) говорит скорее о неправильном построении схемы потока данных, чем о некой задумке автора потока данных. Более подробно и строго об неоднозначностях в поведении потока данных будет идти речь ниже. Стоит заметить, что подобного рода неоднозначности возникают и в процессе определения сетей процессов Кана, которые решаются наложением условий на правила поглощения и испускания, а также на класс функции процесса. Мы же пойдем другим путем, накладывая ограничения лишь на граф шаблона потока данных, оставляя подпрограммы и правила поглощения каждого шаблона блока без ограничений. В нашем случае правила поглощения состоят в том, что формально блок может поглотить любые доступные данные из любых портов по любым ребрам, но будем расценивать любые возможные неоднозначности как ошибку построения потока данных.

Выше были рассмотрены определения и термины, которые с той или иной точностью присутствуют практически во всех моделях потока дан-

ных. В данной работе мы абстрагируемся от данных, а значит нас интересуют только возможные зависимости наборов входных портов от набора выходных портов, иными словами только качественное поведение блоков. Оказывается, конечный автомат Мили хорошо подходит для описания качественного поведения блоков. Заметим, что в данной работе используется не стандартное определение автомата Мили, а его недетерминированный аналог. Дадим теперь строгое определение конечного автомата Мили.

Определение 5. *Конечный автомат Мили* — кортеж $FA = (S, s_0, \Sigma, \Lambda, E)$, где:

- S — конечное множество состояний;
- $s_0 \in S$ — начальное состояние;
- Σ — входной алфавит;
- Λ — выходной алфавит;
- $E \subseteq S \times \Sigma \times \Lambda \times S$ — отношение возможных переходов.

Для удобства будем рассматривать функцию конечного автомата Мили $FA = (S, \cdot, \Sigma, \Lambda, E)$: $FA(p, \sigma) = \{(\lambda, q, \sigma') | (p, \sigma', \lambda, q) \in E, \sigma' \subseteq \sigma\}$. Стоит обратить внимание на нестандартную форму данной функции. Далее символами в алфавитах Σ и Λ будут множества, поэтому функция по текущему состоянию и входному множеству возвращает множество троек вида: выходной символ-множество, новое состояние и поглощенное множество.

Конечный автомат Мили выполняет преобразование последовательности входных символов из алфавита Σ в последовательность символов в алфавите Λ . Вернемся теперь к определению шаблона атомарного блока. В данной работе мы абстрагируемся от конкретного алгоритма преобразования входных данных в выходные и преобразования внутреннего состояния блока. Вместо этого мы будем описывать лишь качественное поведение алгоритма автоматом Мили FA , в котором входной алфавит $\Sigma = 2^I \setminus \emptyset$ (запрещается запуск по пустому набору входных портов), а выходной алфавит $\Lambda = 2^O$, иными словами ребро $e = (u, is, os, v)$ автомата FA описывает возможный переход из состояния u в состояние v , при считывании данных из портов is и выписывании данных в порты os (именно поэтому функция FA имеет странный на первый взгляд вид). На практике подпрограмма реализующая логику атомарного блока может иметь значительное количество внутренних состояний, однако, для наших целей важны лишь группы таких внутренних состояний, которые описывают возможное поведение на определенном наборе данных из входных портов. Более того практически во всех системах построения и запуска потока данных подпрограмма описывающая блок имеет детерминированное поведение. Неизбежный недетерминизм в автомате Мили, описывающий блок, возникает из-за исключения из рассмотрения самих данных, так как при разных значениях на одном и том же наборе портов из одного состояния, блок может перейти в различные конечные состояния, даже выписав при этом данные в один и тот же набор портов.

Пример 1. На рисунке 1b изображен автомат Мили для шаблона атомарного блока *Loor*, описывающего цикл. Поясним эту схему. Изначально

блок с шаблоном *Loop* находится в состоянии *initial*. Далее он может считать из порта *xs* начальные параметры цикла (например, некую коллекцию данных). Если условие выхода из цикла изначально выполнено (например, пустая коллекция), то блок переходит в прежнее состояние выписав в порт *fs*. Иначе, блок отправляет данные по порту *x* переходя в состояние *non_trivial*. В состоянии *non_trivial* блок имеет не тривиальное внутреннее состояние, например, остаток коллекции. В этом состоянии блок ожидает поступление преобразованных некой внешней функцией данных на порт *f*. И в этом случае существует два варианта перехода, в зависимости от выполнения условия выхода из цикла: $(non_trivial, \{f\}, \{x\}, non_trivial)$ и $(non_trivial, \{f\}, \{fs\}, trivial)$ для случаев продолжения цикла и выхода из него соответственно.

Конечный автомат Мили может быть также определен для составного блока. Единственное отличие от автомата Мили для атомарного блока состоит в том, что зная все автоматы внутренних блоков, можно вычислить автомат составного блока как будет показано ниже.

1.4.1 Алгоритм работы потока данных ***Need review***

Теперь опишем алгоритм вычисления потока данных, представленного в виде составного блока c со схемой $C = (\hat{B}_c = B_c \cup \{SOURCE, STOCK\}, \hat{E}_c = E_c \cup E_c^I \cup E_c^O)$ в представлении графа. Мы будем рассматривать вычисление потока данных в предположении, что время работы блока может быть абсолютно любым (но конечным) при любых условиях, что полностью соответствует реальным потокам данных. В таком случае нас прежде всего будет интересовать всевозможные поведения потока данных с точки зрения относительных расположений времен запуска и завершения любого блока. Введем модельное время $t \in \mathbb{Z}_+$. Каждое множество, связанное с состоянием системы в момент времени t , снабдим верхним индексом, например, $\omega^t \in \Omega_c$.

Определение 6. Определим некоторые вспомогательные функции.

$$\begin{aligned} O_v(\omega) &= \{v^O | (v, v^O, \cdot, \cdot) \in \omega\}, \\ I_v(\omega) &= \{v^I | (\cdot, \cdot, v, v^I) \in \omega\}, \\ \Delta_{O,v} &= \{(v, v^O, \cdot, \cdot) \in E_c | v^O \in O\}, \\ \Delta_{I,v}(\omega) &= \{\tilde{\omega}_{I,v} \subseteq \omega\}, \forall \tilde{\omega}_I \forall v^I \in I \exists! e \in \omega_I : e = (\cdot, \cdot, v, v^I) \end{aligned}$$

Последняя функция возвращает множество вариантов поглощения по набору входных портов I блока v .

Определение 7. Фронтон $\Psi_c(\omega, s)$ активной волны $\omega = \{(u, u^O, v, v^I)\} \in \Omega_c$ в состоянии $(s = (s_{u_1}, s_{u_2}, \dots, s_{u_m}), \omega, \cdot)$ назовем подмножество блоков $\Psi_c(\omega, s) \subseteq B_c$, такое что:

$$\Psi_c(\omega, s) = \{v \in B | FA_v(s_v, I_v(\omega)) \neq \emptyset\}$$

Иными словами, фронт активной волны в заданном состоянии — множество блоков, которые могут быть запущены на следующем шаге.

Определение 8. Семейство $\pi_v^t \in E_v$, $FA_v = (\cdot, \cdot, E_v, \cdot)$ показывает текущий переход, который совершает блок v в момент времени t .

Заметим, что значение π_v^t имеет смысл только для работающих на шаге t блоков.

Определение 9. Условием излучения назовем $\zeta(\omega^t, \omega^O) \Leftrightarrow (\omega^O \cap \omega^t = \emptyset)$. Также определим семейства ζ_v^t :

$$\zeta^t = \{v \in B_c | \text{zeta}(\omega^t, \omega_v^O)\}$$

$$\omega_v^O = \{e \in E_c | e = (v, v^O, \cdot, \cdot), v^O \in O_v^t\}, (\cdot, O_v^t, \cdot, \cdot) \in \pi_v^t$$

Условие излучение звучит довольно просто: блок может излучить, если ребра соединенные с выходным набором портов свободны.

Пусть $\phi^t \subseteq \vartheta^t$ — множество блоков, которые закончат работу после шага t . Для любого времени это множество может быть произвольным подмножеством ϑ^t . Однако для любого семейства этих множеств должно выполняться условие конечной работы блока: $\forall t \forall b \in \theta^t \exists \tau > t : b \in \phi^\tau$.

Так же для удобства введем обозначение для любого множества A_c блоков из $B_c - A_c(v) \Leftrightarrow v \in A_c, v \in B_c$.

Теперь выпишем алгоритм преобразования потока данных в виде условий на переходы каждого блока из одного множества в другое. В каждый момент времени блок v может находиться в одном из трех состояний (множеств): свободен ($v \in \xi^t$), работает ($v \in \theta^t$), заблокирован ($v \in \chi^t$). Время работы блоков задается $\phi^t \subseteq \vartheta^t$ — множеством блоков, которые закончат работу после шага t . Для любого времени это множество может быть произвольным подмножеством ϑ^t . Однако для любого семейства этих множеств должно выполняться условие конечной работы блока: $\forall t \forall b \in \theta^t \exists \tau > t : b \in \phi^\tau$. Эти состояние взяты из реальных систем запуска потоков данных и служат для интуитивного понимания модели. В дальнейшем, однако, мы избавимся от этих состояний. Также введем семейства — переход автомата Мили блока, который совершает блок в текущий момент.

$$\xi^t(v) \wedge \neg \Psi_c(\omega^t, s^t)(v) \Rightarrow \begin{cases} \xi^{t+1}(v), \\ s_v^{t+1} = s_v^t, \\ \omega_{I,v}^t = \omega_{O,v}^t = \emptyset \end{cases} \quad (1)$$

$$\xi^t(v) \wedge \Psi_c(\omega^t, s^t)(v) \wedge v \neq \text{STOCK} \Rightarrow \begin{cases} \theta^{t+1}(v), \\ (\cdot, \cdot, I_v^t) \in FA_v(s_v^t, I_v(\omega^t)), \\ \pi_v^{t+1} \in \{(s_v^t, I_v^t, \cdot, s_v^{t+1}) \in E_v\}, \\ \omega_{I,v}^t \in \Delta_{I_v^t, v}(\omega^t), \\ \omega_{O,v}^t = \emptyset \end{cases} \quad (2)$$

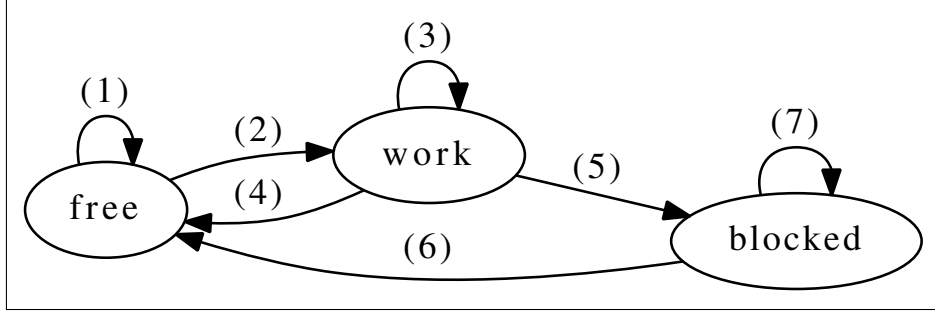


Рис. 3: Схема работы блока

$$\theta^t(v) \wedge \neg \phi^t(v) \Rightarrow \begin{cases} \theta^{t+1}(v), \\ s_v^{t+1} = s_v^t, \\ \pi_v^{t+1} = \pi_v^t, \\ \omega_{I,v}^t = \omega_{O,v}^t = \emptyset \end{cases} \quad (3)$$

$$\theta^t(v) \wedge \phi^t(v) \wedge \zeta^t(v) \Rightarrow \begin{cases} \xi^{t+1}(v), \\ (s_v^t, \cdot, O_v^t, s_v^{t+1}) = \pi_v^t, \\ \omega_{O,v}^t = \Delta_{O_v^t, v}, \\ \omega_{I,v}^t = \emptyset \end{cases} \quad (4)$$

$$\theta^t(v) \wedge \phi^t(v) \wedge \neg \zeta^t(v) \Rightarrow \begin{cases} \chi^{t+1}(v), \\ \pi_v^{t+1} = \pi_v^t, \\ \omega_{O,v}^t = \omega_{I,v}^t = \emptyset \end{cases} \quad (5)$$

$$\chi^t(v) \wedge \zeta^t(v) \Rightarrow \begin{cases} \xi^{t+1}(v), \\ (s_v^t, \cdot, O_v^t, s_v^{t+1}) = \pi_v^t, \\ \omega_{O,v}^t = \Delta_{O_v^t, v}, \\ \omega_{I,v}^t = \emptyset \end{cases} \quad (6)$$

$$\chi^t(v) \wedge \neg \zeta^t(v) \Rightarrow \begin{cases} \chi^{t+1}(v), \\ s_v^{t+1} = s_v^t, \\ \pi_v^{t+1} = \pi_v^t, \\ \omega_{O,v}^t = \omega_{I,v}^t = \emptyset \end{cases} \quad (7)$$

$$\omega^{t+1} = \omega^t \setminus (\cup_{v \in B_c} \omega_{I,v}^t) \cup (\cup_{v \in B_c} \omega_{O,v}^t), \quad (8)$$

$$s^{t+1} = (s_{v_1}^t, \dots, s_{v_m}^t) \quad (9)$$

Работу каждого блока можно схематично изобразить в виде автомата, см. рисунок 3.

Поясним эти правила. Каждое правило (1) — (7) описывает условие перехода в конечном автомате на рисунке 3.

Правила (1) и (2) говорят о поведении свободного блока — если блок может запуститься из его текущего состояния s_v^t (то есть хватает данных

на портах для какого-либо перехода), то он поглощает данные $\omega_{I,v}^t$ и начинает некий переход π_v^{t+1} из текущего состояния, иначе остается свободным дальше.

Тройка правил (3), (4) и (5) описывают поведение работающего блока v . Если множество ϕ^t не содержит рассматриваемый блок, он продолжает работу. Иначе, возможны два варианта: либо блок может завершиться и испустить данные $\zeta^t(v)$ (то есть ребра $\Delta_{O_v^t,v}$, связанные с испускаемыми портами O_v^t , свободны), то блок благополучно завершает работу $\xi^{t+1}(v)$ и испускает данные $(\omega_{O,v}^t)$, либо блок переходит в состояние блокировки до момента, когда требуемые ребра освободятся, которое описывается правилами (6) и (7).

Заметим, что мы специально различаем состояния блокировки и работы, хотя можно было бы считать, что заблокированный блок просто работает дополнительное время. Состояние блокировки отличается тем, что блок в этом состоянии не требует вычислительных ресурсов, что будет важно в дальнейшем.

Заметим, что в правилах поведения потока данных сразу видны все неопределенности, о которых говорилось ранее — перед запуском блока, определяется переход, который будет совершать блок, вместе с набором поглощаемых портов и ребра с которых эти данные будут поглощены. Подробнее мы будем рассматривать эти неопределенности далее.

1.4.2 Начало и завершение работы потока данных

Если рассматривать представление потока данных в виде графа, то начало и завершение потока данных связано с дополнительными или 'фиктивными' блоками SOURCE и STOCK: их задачи сформировать начальную и поглотить конечную активные волны, соответственно. Исходя из правил, можно записать начальные условия:

$$\omega^0 = \emptyset, \quad (10)$$

$$s_v^0 = s_{0,v}, \forall v \in B_c, \quad (11)$$

$$\theta^0 = \{\text{SOURCE}\}, \quad (12)$$

$$\xi^0 = B_c \setminus \{\text{SOURCE}\}, \quad (13)$$

$$\chi^0 = \emptyset, \quad (14)$$

$$\omega^1 \neq \emptyset \quad (15)$$

Формирование начальной волны заложено в принципе работы блока SOURCE. Так как блоки SOURCE и STOCK являются вспомогательными, то нет смысла рассматривать их автоматы Мили, так как автомат для блока SOURCE не может иметь ни одного перехода из-за отсутствия входных портов, а автомат для блока STOCK будет создавать дополнительные неопределенности при анализе потока данных. Будем считать, что работа этих блоков определяется следующим образом: SOURCE на первом шаге испускает некий непустой набор входных данных — начальную активную волну; блок STOCK поглощает все данные на его портах, если никакие другие блоки больше не могут запуститься. Именно запуск блока STOCK мы будем считать завершением работы всего потока данных, иными словами,

поток данных завершает работу на шаге $t = T_{end}$, если:

$$\theta^{T_{end}} = \emptyset \wedge \Psi_c(\omega^{T_{end}}, s^{T_{end}}) = \{\text{STOCK}\} \quad (16)$$

Завершение работы потока данных связано с понятием траектории запуска потока данных и более подробно будет рассмотрено ниже.

2 Анализ модели *** Need review ***

2.1 Построение конечного автомата Мили по программе блока ***Need review***

Введенная выше модель оперирует с упрощенными описаниями алгоритмов блоков — каждый алгоритм заменяется на описывающий его качественное поведение автомат Мили. Каждое состояние конечного автомата соответствует группам внутренних состояний алгоритма блока. Разбиение на группы можно производить с большой свободой. Сразу же заметим, для дальнейшего анализа будет полезно выделить особое состояние конечного автомата Мили для блока — начальное состояние (далее будет обозначаться словом INITIAL), соответствующее начальным значениям внутренних переменных. Это состояние обладает значительной особенностью, а именно, блок находящийся в нем, фактически не занимает ресурсов памяти вычислительного узла, поэтому может быть просто перенесен на другой вычислительный узел без дополнительных затрат на копирование внутреннего состояния с одного вычислительного узла на другой. Процедуры перемещения состояния алгоритма блока на другой вычислительный узел играют важную роль в управлении запуском потока данных в распределенной среде (более подробно запуск потока данных в распределенной среде будет рассмотрен ниже), так как ожидающий данных блок будет заблокирован (или по, крайней мере, может быть дополнительно ограничен в вычислительных ресурсах), если на вычислительном узле на момент поступления данных первому блоку выполняется алгоритм другого блока, но может быть инициализирован на свободном (если таковые имеются) вычислительном узле, если блок находится в начальном состоянии.

Построение автомата Мили для алгоритма блока — неоднозначная задача, требующая понимания качественного поведения алгоритма, а значит участия человека. В данной работе подробно не рассматриваются алгоритмы (которые, по мнению автора, будут бесполезны на практике в силу огромных вычислительных затрат) и методики построения автоматов Мили для алгоритмов блоков. Ограничимся лишь рассмотрением показательного примера с шаблоном блока *For-Loop*. В зависимости от реализации, блок *For-Loop* принимает на вход данные о количестве необходимых операций, для определенности будем рассматривать шаблон блока, реализующий аналог функции *map* популярной в функциональных языках программирования, которая поэлементно применяет переданную ей в качестве аргумента функцию к списку значений и выдает список результатов. Типичный пример составного блока с блоком шаблона *For — Loop* показан на рисунке 2а (сам блок шаблона *For-Loop* именуется *map*), а автомат Мили шаблона блока *For-Loop* показан на рисунке 1b. Сам блок шаблона *For-Loop* требует внешнего блока-функции, который подключен к портам x и f . Между вызова-

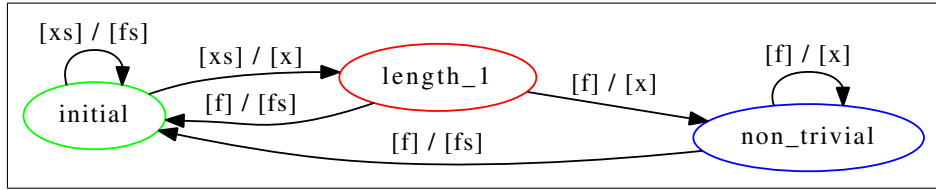


Рис. 4: Избыточный конечный автомат Мили для шаблона блока *For-Loop*

ми (то есть испусканием очередного элемента списка) блока-функции, блок шаблона *For-Loop* имеет внутреннее состояние, в котором хранит, как минимум, не обработанную часть списка и частичный результат преобразования. Поэтому в конечном автомате Мили шаблона блока *For-Loop* имеются два состояния: *initial* — начальное состояние ожидания данных, *non_trivial* — состояние, в котором результирующий список находится в стадии формирования.

Фактически, для любого алгоритма можно построить автомат Мили с не более, чем двумя состояниями: начальным (или тривиальным) и не тривиальным (с отличным от начального внутренним состоянием) состояниями, однако, в некоторых случаях автомат Мили будет вносить дополнительную неопределенность в поведение блока, не связанную с неопределенностью входных данных при анализе. Возможен и обратный случай — в предыдущем примере с шаблоном блока *For-Loop* можно ввести отдельное состояние, отвечающее за списки длины 1. В данном случае автомат Мили будет выглядеть, как показано на рисунке 4, дополнительное состояние именуется *length_1*. Это дополнительное состояние, конечно, несет новую информацию о текущем состоянии блока, но она абсолютно бесполезна для анализа схемы потока данных.

2.2 Эквивалентность траекторий запуска потока данных ***Need review***

Рассмотрим подробнее правила работы потока данных. В правилах помимо неопределенностей в выборе вариантов поведения, присутствует множество ϕ^t , задаваемое извне. Фактически это множество определяет время работы каждого запущенного блока и введено для возможности описания любой возможной конфигурации времен работы алгоритмов блоков. Однако при анализе потока данных эти времена неизвестны и здесь не будут рассматриваться их оценки. Для анализа большее значения играет не сами времена работы блоков, а зависимости типа "запуск блока *A* был вызван завершением работы блоков B_1, B_2, \dots, B_l ". В дальнейшем, для выяснения всевозможных поведений потока данных в терминах этих зависимостей, мы будем эмулировать запуск потока данных по правилам (1)-(9), и, в частности, положим времена работы всех блоков равными единице. Модельное время в данном случае отражает лишь один из возможных случаев запуска потока данных и служит только для удобства понимания. Вначале определим понятие траектории запуска потока данных и введем на множестве траекторий запуска отношение эквивалентности.

Определение 10. Траекторию запуска потока данных $\Lambda = \{\Lambda^t\}_{t=0}^{\infty}$ мы определим как семейство всех множеств упомянутых в правилах работы

потока данных (1)-(9) и описывающих состояние потока данных, то есть $\Lambda^t = (\omega^t, \{\omega_{I,v}^t | v \in B_c\}, \{\omega_{O,v}^t | v \in B_c\}, \pi^t, s^t, \xi^t, \theta^t, \chi^t$.

Введем также определения конечной и успешной траекторий и времени работы потока данных на этой траектории.

Определение 11. Траектория запуска потока данных $\Lambda = \{\Lambda^t\}_{t=0}^\infty$ называется конечной, если $\exists T_{end} : \forall t \geq T_{end} : \theta^t = \emptyset$. Минимальное T_{end} называется временем работы потока данных.

Определение 12. Траектория запуска потока данных $\Lambda = \{\Lambda^t\}_{t=0}^\infty$ называется успешной, если $\exists T_{end} : \forall t \geq T_{end} : \xi^t = B_c \wedge \omega^t = \emptyset$.

Понятно, что минимальные T_{end} из определений 11 и 12 совпадают (если оба существуют), и более того, равны (с точностью до времени работы блока СТОК, которое, для удобства положим равным 0) T_{end} в (16), так как при отсутствии работающих блоков фронт активной волны ω^t не может меняться (в силу правила работы потока данных (8)), а значит правило (6) не может быть выполнено для заблокированных блоков, так как оно полностью определяется фронтом активной волны. В дальнейшем мы будем предполагать, что все траектории рассматриваемого потока данных являются конечными.

Множества π_v^t , $\omega_{I,v}^t$ и $\omega_{O,v}^t$ полностью описывают отношение эквивалентности между траекториями.

Определение 13. Для каждой конечной траектории запуска множества π_v^t , I_v^t порождают граф причинности $G = (S, E)$. Для каждого блока прономеруем все времена его запуска t_v^n . Тогда G можно определить как:

- $S = \{(v, n, \pi_v^{t_v^n})\}$,
- из $s_1 = (v, \cdot, \pi_1)$ в $s_2 = (v, \cdot, \pi_2)$ есть ребро тогда и только тогда, когда часть поглощенной $\omega_{I,v}^t$ в результате перехода π_2 была испущена в результате перехода π_1 ; при этом ребро дополнительно помечается множеством соответствующих входных портов блока v .

Отображение $G_\Lambda = G(\Lambda)$ индуцирует классы эквивалентности траекторий запуска потока данных, иными словами $\Lambda_1 \sim \Lambda_2 \Leftrightarrow G(\Lambda_1) = G(\Lambda_2)$. Отметим, что данное отношение эквивалентности имеет смысл только для успешных траекторий, так как в нем не учитываются остаточные активные волны и заблокированные блоки. Наличие неуспешных траекторий потока данных мы будем рассматривать как ошибку проектирования схемы потока данных. Отметим, что это правило полностью соответствует практике.

Каждый граф причинности качественно описывает запуск потока данных.

Пример 2. Рассмотрим один показательный пример нахождения графов причинности для потока данных, содержащего цикл (схема потока данных изображена на рисунке 2а). Для простоты, будем считать, что автомат Мили блока f состоит из одного состояния и одного перехода $\pi_f = (INITIAL, \{x\}, \{f\}, INITIAL)$. Наличие ребра из s_1 в s_2 в графе причинности будем обозначать как $s_1 \rightsquigarrow s_2$.

Начальная вершина всех графов причинности — $s_0 = (SOURCE, 0, \pi_{SOURCE})$. Далее возможны варианты: либо начальная вершина вызовет переход $\pi_{MAP,1} = (INITIAL, \{xs\}, \{fs\}, INITIAL)$ (то есть переход по пустому списку данных), либо $\pi_{MAP,2} = (INITIAL, \{xs\}, \{x\}, NON-TRIVIAL)$. Последний в свою очередь вызовет переход π_f , который, в свою очередь, может вызвать переходы $\pi_{MAP,3} = (NON-TRIVIAL, \{f\}, \{x\}, NON-TRIVIAL)$ и $\pi_{MAP,4} = (NON-TRIVIAL, \{f\}, \{xs\}, INITIAL)$, и так далее. Конечная вершина графа — $s_{end} = (STOCK, T_{end}, \pi_{STOCK})$. Получаем семейство графов причинности:

$$\begin{aligned} s_0 &\rightsquigarrow (MAP, t_1, \pi_{MAP,1}) \rightsquigarrow s_{end}, \\ s_0 &\rightsquigarrow (MAP, t_1, \pi_{MAP,2}) \rightsquigarrow (f, t_2, \pi_f) \rightsquigarrow (MAP, t_3, \pi_{MAP,4}) \rightsquigarrow s_{end}, \\ s_0 &\rightsquigarrow (MAP, t_1, \pi_{MAP,2}) \rightsquigarrow (f, t_2, \pi_f) \rightsquigarrow (MAP, t_3, \pi_{MAP,3}) \rightsquigarrow \\ &\rightsquigarrow (f, t_4, \pi_f) \rightsquigarrow \dots \rightsquigarrow (MAP, t_n, \pi_{MAP,4}) \rightsquigarrow s_{end} \end{aligned}$$

Этот пример демонстрирует необходимость следующего предположения: будем считать, что любая циклическая зависимость конечна (то есть не будем рассматривать не конечные траектории, возникающие из-за этих зависимостей).

2.3 Неопределенности в запуске потока данных ***Need review***

Как было упомянуто выше, в анализе схемы потока данных неизбежно возникают два рода неопределенностей: первый вызван отсутствием знания входных данных при анализе схемы, неопределенности второго рода связаны с самой схемой потока данных, которая может допускать такие неоднозначности как, например, несколько вариантов поглощения данных. Неоднозначности первого рода неизбежны, и поэтому следует обрабатывать все варианты порожденные ими. Все эти неоднозначности второго рода возникают как следствие состояний гонки (race condition), которых традиционно стараются избегать, так как результат выполнения всего потока данных становится зависим от времен работы блоков, которые, в свою очередь, могут зависеть от данных и параметров вычислительных узлов, что говорит скорее о неправильном построении схемы потока данных.

Сформулируем теперь критерии неопределенностей первого и второго рода.

Любые неопределенности связаны с правилом запуска потока данных (2) — правило запуска блока, что видно из правил работы потока данных, так как, только это правило содержит включения, вместо строгих равенств.

Определение 14. Будем говорить, что состояние s автомата Мили $FA_v = (S_v, \cdot, \cdot, \cdot, E_v)$ блока v допускает неопределенность первого рода, тогда и только тогда, когда: $\exists I_v : |\{e \in E_v | e = (s, I_v, \cdot, \cdot)\}| > 1$, или иными словами, существует набор входных портов I_v , такой что, из состояния s возможны несколько переходов по входным портам I_v .

Например, в автомате Мили шаблона блока *For-Loop* (рисунок 1b) каждое состояние допускает неопределенность первого рода. Причины возникновения этих неопределенностей были описаны выше.

Определение 15. Блок $v \in B_c$ в состоянии Λ^t траектории запуска потока данных на шаге t допускает неопределенность второго рода, если: $|\omega_{I,v}^t| > 1$, где $\omega_{I,v}^t$ определена в правиле (2) и содержится в кортеже Λ^t .

Определение 16. *Корректным потоком данных* будем называть поток данных, любая возможная траектория которого не допускает неопределенностей второго рода.

Заметим, что по автомату Мили можно сразу предсказать возможные неопределенности второго рода, например, если автомат Мили содержит переходы $\pi = (s, I, \cdot, \cdot)$ и $\pi' = (s, I', \cdot, \cdot)$, где $I \subseteq I'$: в этом случае переход по π' автоматически означает неопределенность второго рода, соответственно, наличие таких переходов автоматически означает неправильность построения автомата Мили или алгоритма блока.

Введем вспомогательное определение набора последовательностей времен работы блоков.

Определение 17. Зафиксируем набор последовательностей времен работы $\tau = \{\{\tau_v^n\}_{n=0}^\infty | v \in B_c\}$. Траектория Λ согласуется с набором последовательностей времен запуска, если время работы n -го запуска блока v $t_v^n = \tau_v^n$.

Теорема 1. В работе потока данных не существует других неопределенностей, кроме неопределенностей первого и второго рода, то есть, если зафиксировать набор последовательностей времен работы блоков τ , то любое ветвление префиксного дерева траекторий, согласующихся с τ , описывается только неопределенностями первого и второго рода.

Доказательство. Доказательство напрямую следует из факта, что любые различия в траекториях запуска потока данных при фиксированных временах работы блоков определяются правилом (2) работы потока данных. \square

Среди неопределенностей второго рода следует выделить важный случай, уже упоминавшийся выше — состояние гонки.

Определение 18. Поток данных допускает состояние гонки, если существует некая последовательность времен запуска τ и класс эквивалентности \mathbb{K} траекторий запуска потока данных, такие что траектории удовлетворяющие τ не лежат в классе \mathbb{K} .

Поясним это определение. Из определения сразу же следует, что в отсутствии состояния гонки, каждый класс эквивалентности для любого набора последовательностей времен запуска содержит траекторию, согласующуюся с ней, что означает то, что результат не зависит от времен работы блоков (что соответствует классическим определениям).

2.4 Волновой алгоритм

Решение задачи нахождения классов эквивалентности траекторий запуска позволяет существенно упростить анализ потока данных, так как переход к графам причинности позволяет избавиться от модельного времени для корректных потоков данных.

Однако, даже для корректного потока данных, классов эквивалентности может быть бесконечное количество. Это утверждение было проиллюстрировано в примере 2. Одним из возможных решений этой проблемы — нахождение циклов и их замена на соответствующий составной блок. Однако, это не всегда удастся, например, из-за того, что 'тело' цикла может иметь связи с самим блоком цикла. Однако, если замена циклов удалась, то поток данных будет преобразован к ациклическому графу, работу которого заметно проще анализировать. Однако в данной работе мы пойдем другим путем.

Мы будем находить

2.5 Работа волнового алгоритма в графе переходов