Синхронизация в распределенных системах

Взаимное исключение

Background

- Synchronization: coordination of actions between processes.
- Processes are usually asynchronous, (operate independent of events in other processes)
- Sometimes need to cooperate/synchronize
  - For mutual exclusion
  - For event ordering (was message x from process P sent before or after message y from process Q?)

Introduction

- Synchronization in centralized systems is primarily accomplished through shared memory
  - Event ordering is clear because all events are timed by the same clock
- Synchronization in distributed systems is harder
  - No shared memory
  - No common clock

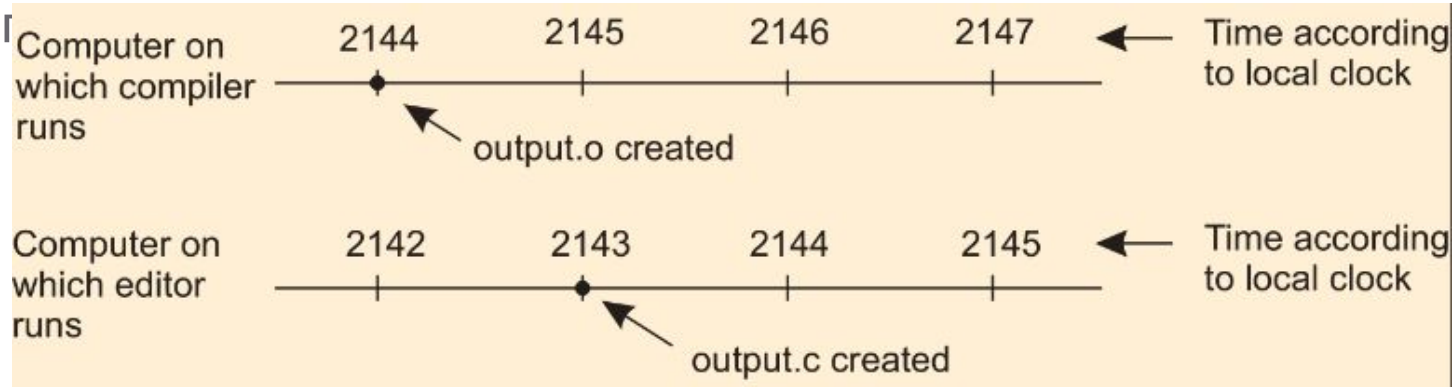
Основные механизмы синхронизации в распределенных системах

- Синхронизация часов
- Логические часы
- Глобальное состояние
- Алгоритмы голосования
- Взаимное исключение
- Распределенные транзакции

Синхронизация времени

Роль системных часов

- Некоторые приложения основываются на реальном порядке событий происходящих в системе
  - Например команда make в OS Unix, которая учитывает время последней модификации файла при перетрансляции модулей



  - Порядок возникновения событий в локальной системе может быть легко учтен на основании меток времени, но в распределенных системах системные часы узлов бут синхронизированы не всегда.

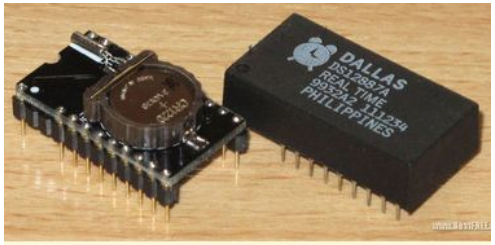
- Возможна ли синхронизация часов в

Физические часы

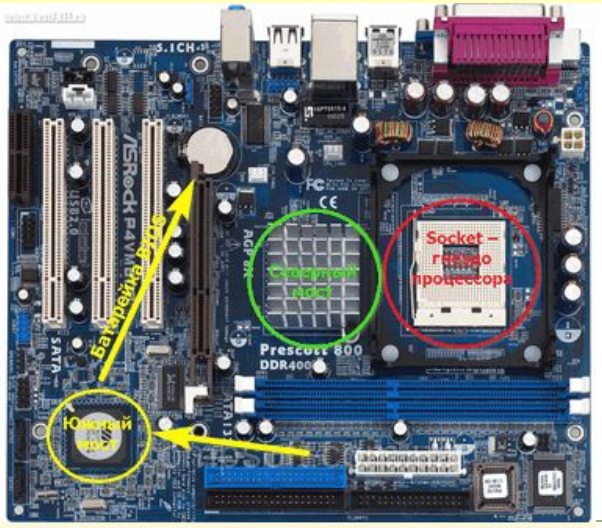
- Солнечные часы (песочные, водяные, огненные и т.п.)
- Механические часы
- Электронные часы
- Системные часы ЭВМ


```
~/user: date "+%Y-%m-%d %H:%M:%S"
2009-01-06 14:18:11
```

FOCS 1, атомные часы в Швейцарии с погрешностью 10^{-15}, то есть не более секунды за 30 миллионов лет

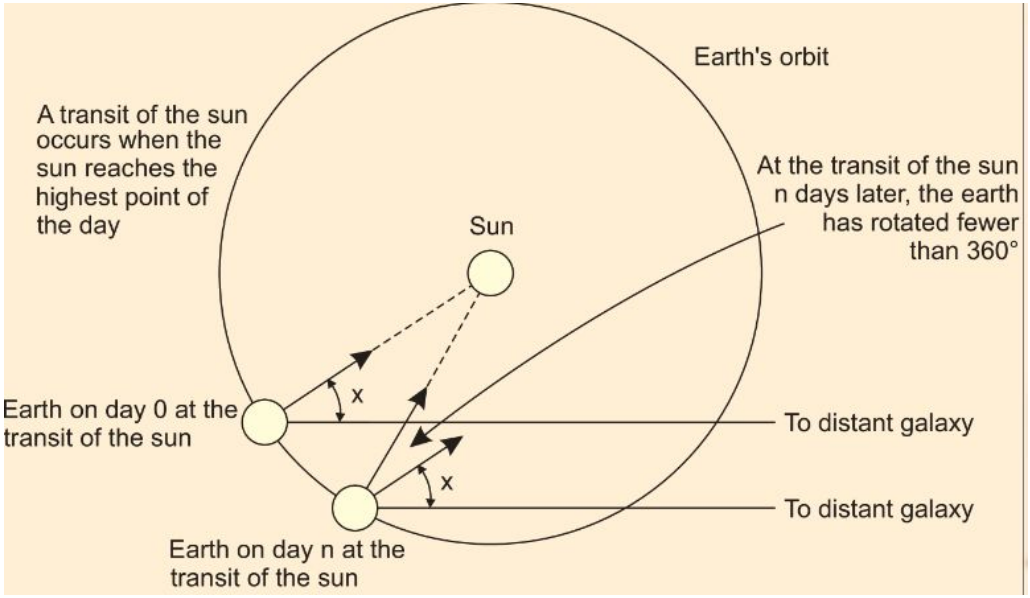


Микросхема часов реального времени Dallas



Южный мост на материнской плате

Солнечная секунда и время по Гринвичу



- Солнечная секунда (solar second) определяется как 1/86 400 солнечного дня.
- Геометрические построения, необходимые для расчета солнечного дня, приведены на рисунке.
- Период обращения замедляется из-за приливного трения и вязкости атмосферы.
- Продолжительность года (время одного оборота вокруг солнца) при этом не изменяется, сутки просто становятся

Среднее время по Гринвичу (англ. *Greenwich Mean Time*, *GMT*), или **гринвичское время** — среднее солнечное время меридиана — среднее солнечное время меридиана, проходящего через прежнее место расположения Гринвичской — среднее солнечное время меридиана, проходящего через прежнее место расположения Гринвичской королевской обсерватории — среднее солнечное время меридиана, проходящего через прежнее место расположения Гринвичской королевской

Архитектура и принцип работы часов реального времени RTC и CMOS памяти.

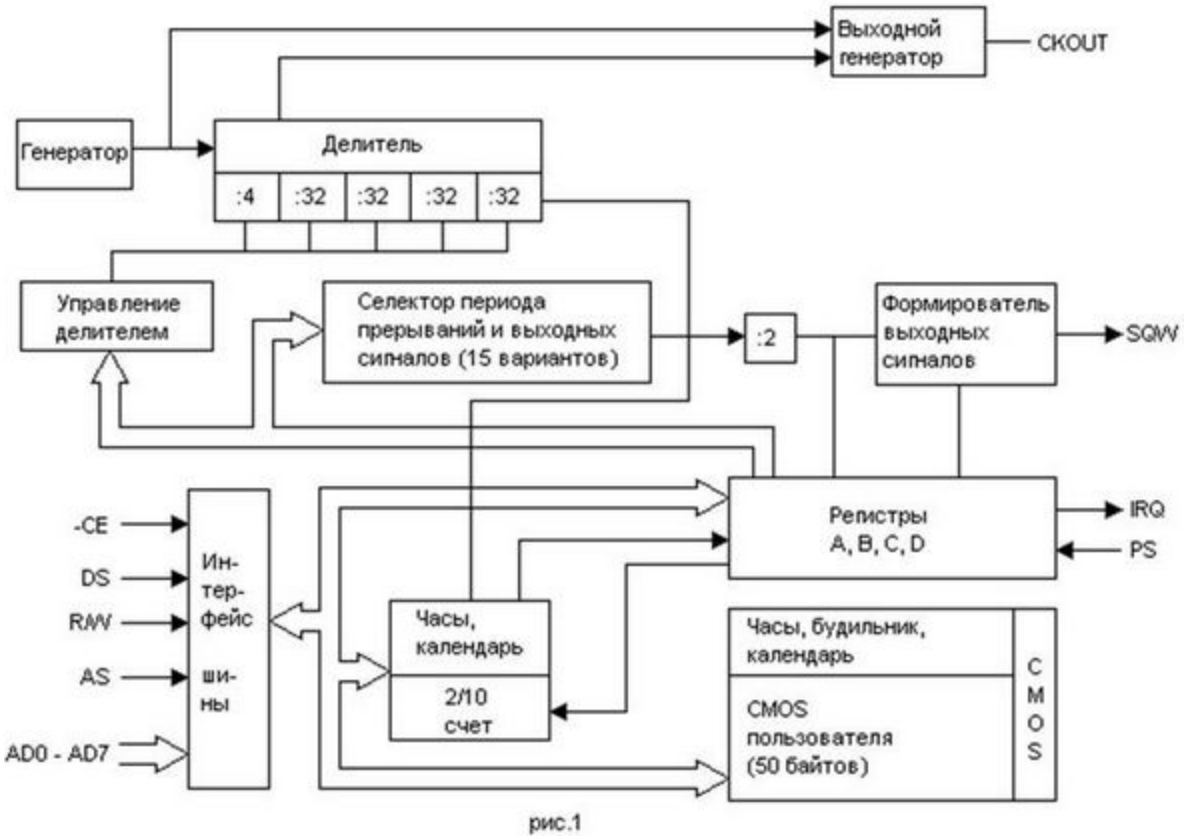


рис.1

Кварцевый генератор имеет начальную погрешность 30 частей от миллиона (parts per million), то есть 32768Гц*30/1000000=±0,98304Гц.

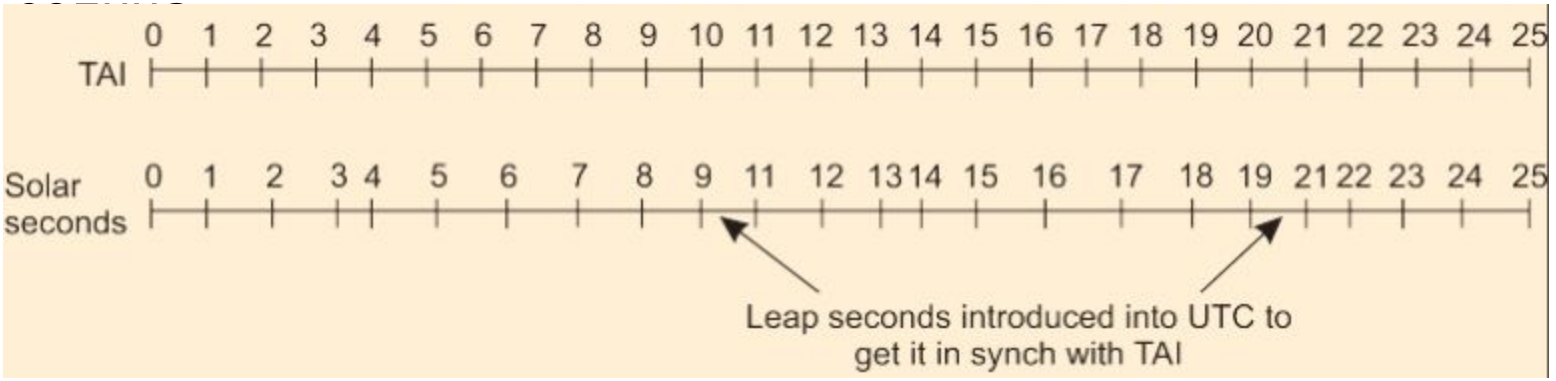
- В состав IBM PC AT входят часы реального времени Real Time Clock (RTC) и 64 байта неразрушающейся оперативной КМОП памяти (CMOS), питающиеся от автономного источника питания.
- При включении ПЭВМ содержимое CMOS анализируется POST, который извлекает из нее конфигурацию системы и текущие дату и время.
- Кварцевый генератор имеет частоту 32768 Гц и

Глобальное время пор атомным часам

- В 1948 году были изобретены атомные часы. Физики определили секунду как время, за которое атом цезия-133 совершит ровно 9 192 631 770 переходов. Выбор числа 9 192 631 770 сделал атомную секунду равной средней солнечной секунде в год ее расчета.
- В настоящее время около 50 лабораторий по всему миру имеют часы на цезии-133. Периодически каждая лаборатория сообщает в международное бюро мер и весов в Париже, сколько времени на их часах. Международное бюро усредняет их результаты и выдает глобальное время по атомным часам {International Atomic Time, TAI). TAI — это среднее время тиков часов на цезии-133, прошедшее с полуночи 1 января 1958 года (начала времен) и деленное на 9 192 631 770.
- Хотя время TAI весьма стабильно, но имеется серьезная проблема: 86 400 с TAI в настоящее время приблизительно на 3 мс меньше среднего солнечного дня (потому что средний солнечный день все время удлиняется).

Универсальное согласованное время (UTC)

- Международное бюро решило эту проблему, используя **потерянные секунды** (leap seconds) всякий раз, когда разница между временем TAI и солнечным временем возрастает до **800 мс**. Эта коррекция позволила перейти к системе, основанной на постоянных секундах TAI, в которой, однако, соблюдается соответствие с периодичностью очевидно видимого движения



- Она называется универсальным согласованным временем {Universal Coordinated Time, UTC). UTC — это основа всей системы хранения времени в наши дни. Оно, по существу, заменило старый стандарт — среднее время по Гринвичу {Greenwich mean time), которое основывалось на астрономических наблюдениях и

Источники точного времени UTC

- National Institute of Standard Time, NIST) имеет коротковолновую радиостанцию с позывными WWV из форта Коллинз (Fort Collins), штат Колорадо. Радиостанция WWV широковещательно рассылает короткий импульс в начале каждой секунды UTC. Точность самой радиостанции WWV составляет около ±1 мс, но из-за различных атмосферных флуктуации длина сигнала может меняться, так что на практике точность составляет не более ±10 мс.
- В Англии станция MSF, работающая из Регби (Rugby), район Варвикшир (Warwickshire), предоставляет похожую службу.
- Существуют также станции UTC и в некоторых других странах.
- Некоторые спутники Земли также предоставляют службы UTC. Рабочий спутник геостационарного окружения (Geostationary Environment Operational Satellite - GEOS) может предоставлять время UTC с точностью до 0,5 мс, а некоторые другие — и с более высокой точностью.
- GPS (Global Positioning System) обеспечивает точность до 20-35

Способы синхронизации часов в распределенных системах

- Если одна машина имеет приемник WWV, то задачей является синхронизация с ней всех остальных машин.
- Если приемников WWV нет ни на одной из машин, то каждая из них отсчитывает свое собственное время, то задачей будет по возможности синхронизировать их между собой.
- Для проведения синхронизации было предложено множество алгоритмов.

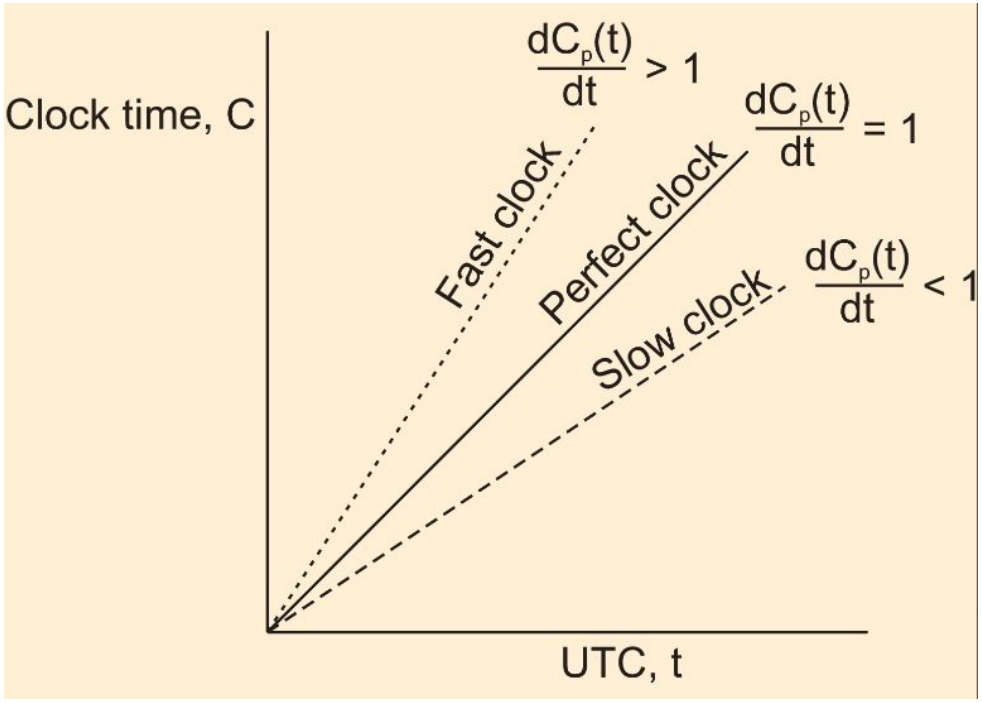
Сдвиг системных часов машин

- Все алгоритмы имеют одну базовую модель системы.
- Считается, что каждая машина имеет таймер, который инициирует прерывание *H раз в секунду.*
- Обозначим значение часов машины p — Cp(t). В идеальном мире мы можем считать, что Cp(t) = t для всех p и всех t. Другими словами, dC/dt — точно единица.
- Теоретически таймер с H= 60 должен генерировать 216 000 тиков в час. На практике относительная ошибка, допустимая в современных микросхемах таймеров, составляет порядка 10^{-5}. Это означает, что конкретная машина может выдать значение в диапазоне от 215 998 до 216 002 тиков в час.
- Пусть имеется константа r (максимальная скорость дрейфа часов системы):

  1-r<=dC/dt<= 1+r

- В этих пределах таймер может считаться работоспособным.

Отстающие и спешащие часы



- Если двое часов уходят от UTC в разные стороны за время A^ после синхронизации, разница между их показаниями может быть не более чем 2r·Δt.
- Если разработчики операционной системы хотят гарантировать, что никакая пара часов не сможет разойтись более чем на Δt, то синхронизация часов должна производиться не реже, чем каждые Δt /2p с.
- Различные алгоритмы отличаются точностью определения момента проведения повторной синхронизации.

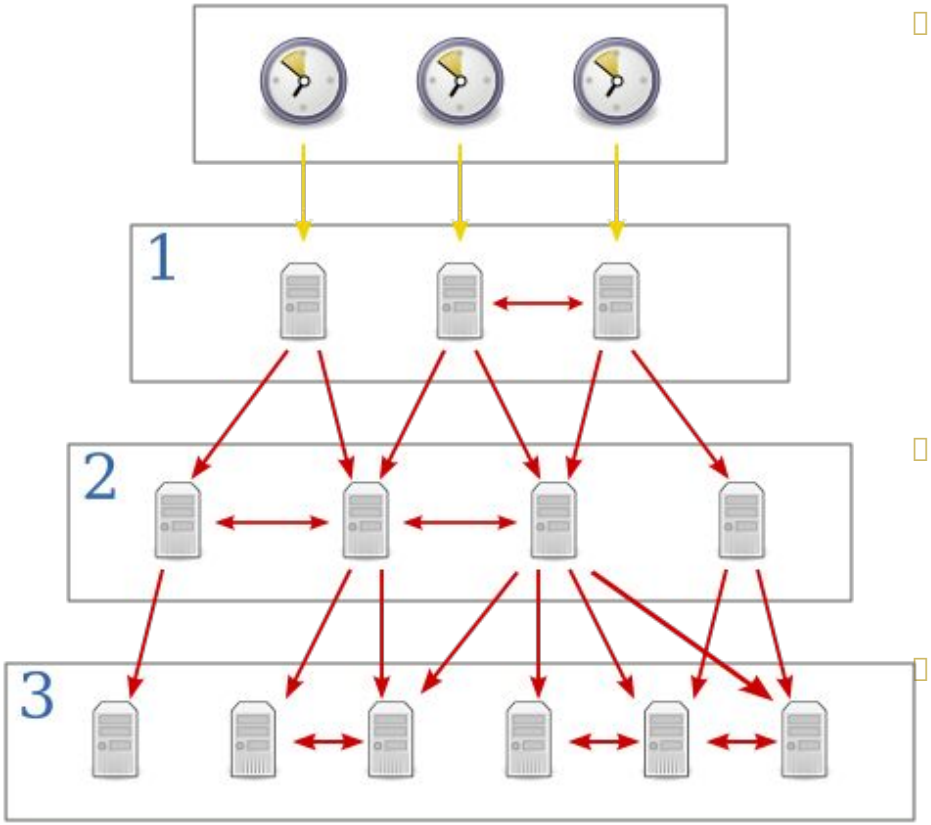
Три философии (цели) синхронизации часов

- Попытаться обеспечить как можно более точную синхронизацию с реальным временем UTC.
- Попытаться обеспечить максимально возможную синхронизацию узлов друг с другом, даже в ущерб синхронизацией с UTC.
- Обеспечить синхронизацию достаточную для обеспечения правильного взаимодействия узлов друг с другом на основе сохранения правильного порядка обмена сообщениями.
  - В этом случае говорят о **логических часах** PC.

Алгоритмы синхронизации часов

- Network Time Protocol (NTP):
  - Цель: обеспечить синхронизацию всех часов по UTC в пределах 1-50мс. Используется в сетях TCP/IP.
  - Для синхронизации используется иерархия пассивных серверов NTP
- Алгоритм Беркли:
  - Цель: обеспечить синхронизацию часов узлов друг с другом (внутренняя синхронизация)
  - Для синхронизации используются активные сервера времени периодически опрашивающие узлы.
- Reference broadcast synchronization (RBS) (Опорная широковещательная синхронизация)
  - Цель: обеспечить синхронизацию часов узлов друг с другом в беспроводной сети

NTP

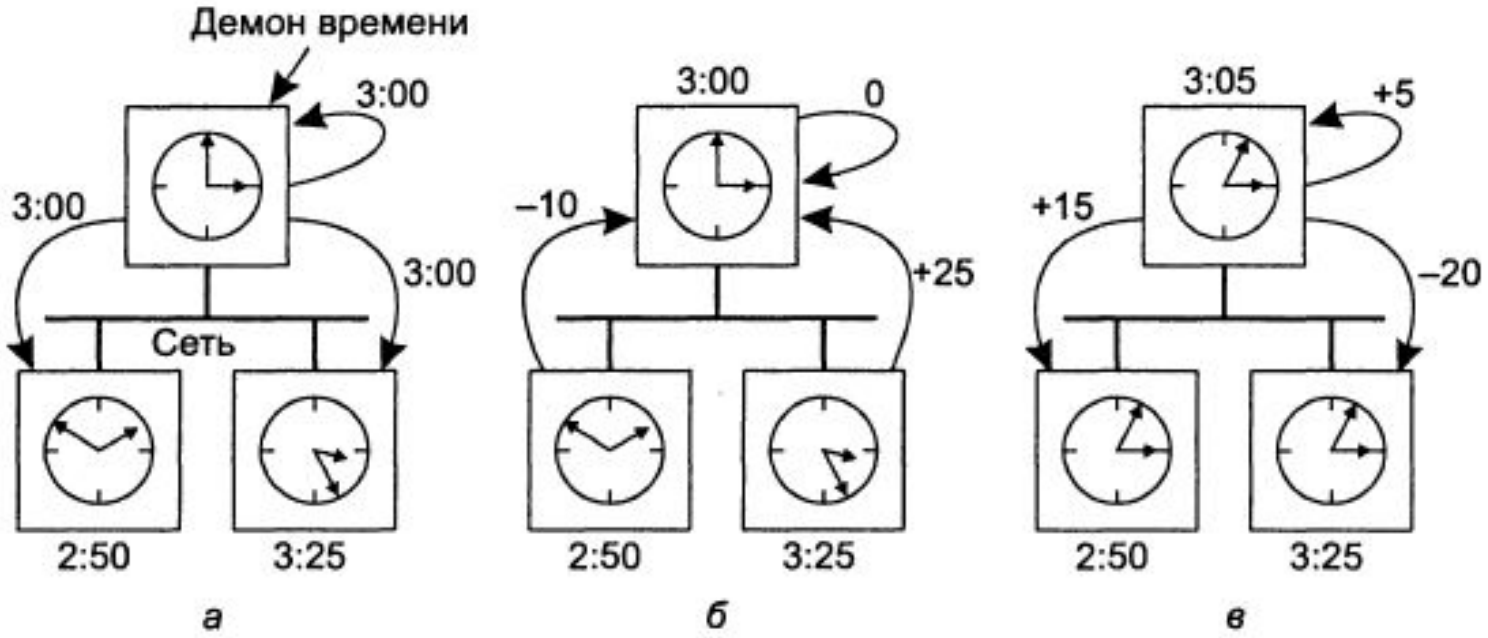


Время представляется в системе NTP 64-битным числом (8 байт), состоящим из 32-битного счётчика секунд и 32-битного счётчика долей секунды, позволяя передавать время в диапазоне 2^{32} секунд, с теоретической точностью 2^{-32} секунды.

- NTP-серверы работают в иерархической сети, каждый уровень иерархии называется ярусом (stratum). Ярус 0 представлен эталонными часами. За эталон берется сигнал GPS (Global Positioning System) или службы ACTS (Automated Computer Time Service). На нулевом ярусе NTP-серверы не работают.
- NTP-серверы яруса 1 получают данные о времени от эталонных часов. NTP-серверы яруса 2 синхронизируются с серверами яруса 1. Всего может быть до 15 ярусов.
- NTP-серверы и NTP-клиенты получают данные о времени от серверов яруса 1, хотя на практике NTP-клиентам лучше не делать этого, поскольку тысячи индивидуальных клиентских запросов окажутся слишком большой нагрузкой для серверов яруса 1. Лучше настроить локальный NTP-сервер, который ваши клиенты будут использовать для получения информации о времени.

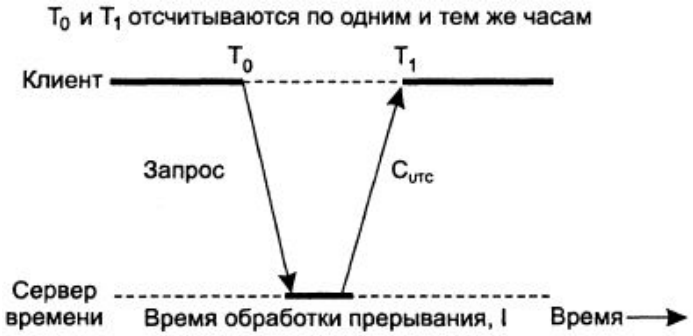
Алгоритм Беркли

- Демон времени (исполняющийся процесс на сервере времени) запрашивает у всех остальных машин значения их часов (а).
- Демон получает ответы машин (б).
- Демон времени сообщает всем, как следует подвести их часы (в)



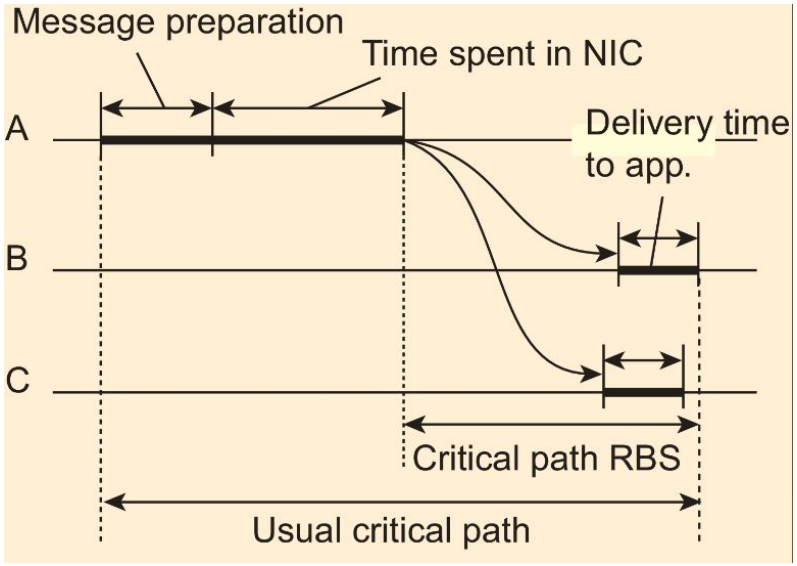
Алгоритм Кристиана

- Машину с приемником WWV назовем *сервером времени.*
- Периодически, гарантировано не реже, чем каждые 2r·Δt с, каждая машина посылает серверу времени сообщение, запрашивая текущее время.



- Когда отправитель получает ответ, он может просто выставить свои часы в значение C_{UTC} . Однако такой алгоритм имеет две проблемы:
  - Главную – время идет только вперед .
  - Второстепенную – ответное сообщение поступает с не нулевой задержкой.

Задержка при передачи значения времени по сети



- По Кристиану, метод решения проблемы состоит в измерении величины задержки передачи по сети.
- Для повышения точности Кристиан предложил производить не одно измерение, а серию. Все измерения, в которых разность T_1-T_0 *превосходит некоторое* пороговое значение, отбрасываются как ставшие жертвами перегруженной сети, а потому недостоверные. Оценка делается по оставшимся замерам, которые могут быть усреднены для получения наилучшего значения.

Множественные внешние источники точного времени

- Для систем, которым необходима особо точная синхронизация по UTC, можно предложить использование нескольких приемников WWV, GEOS или других источников UTC.
- Однако из-за врожденной неточности самих источников времени и флуктуации на пути сигнала лучшее, что могут сделать операционные системы, — это установить интервал, в который попадает UTC. В основном различные источники точного времени будут порождать различные диапазоны, и машины, к которым они присоединены, должны прийти к какому-то общему соглашению.
- Чтобы достичь этого соглашения, каждый процессор с источником UTC может периодически делать широковещательную рассылку своих данных, например, точно в начале каждой минуты по UTC.
- Но печально то, что задержка между посылкой и приемом будет зависеть от длины кабеля и числа маршрутизаторов, через которые должен будет пройти пакет. Эти значения различны для каждой пары (источник UTC, процессор). Будут также играть свою роль и другие факторы, так что точной синхронизации добиться не удастся.

Использование синхронизированных часов

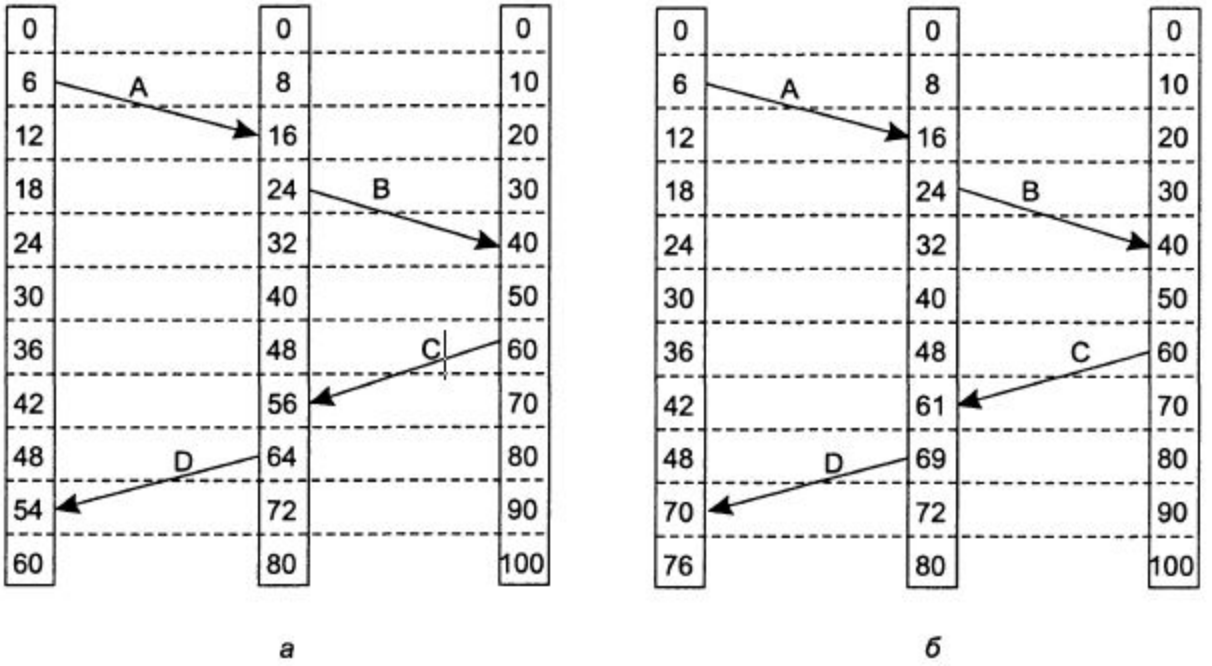
- Благодаря новым технологиям сегодня можно синхронизировать миллионы системных часов с точностью до нескольких миллисекунд по UTC.
- Традиционный подход состоит в том, что каждому сообщению приписывается уникальный номер сообщения, а каждый сервер сохраняет все номера сообщений, которые он принял, чтобы можно было отличить новое сообщение от повторной посылки. Проблема этого алгоритма состоит в том, что при сбое и перезагрузке сервера он теряет эту таблицу номеров сообщений.
- Раньше использовался метод который состоит в сохранении на диске таблицы меток времени и идентификаторов связи.
- Сегодня при загрузке системы выполняется запрос к серверу NTP и синхронизация восстанавливается автоматически.

Логические часы

- Для работы программы make, например, достаточно, чтобы все машины считали, что сейчас 10:00, даже если на самом деле сейчас 10:02. Так, для некоторого класса алгоритмов подобная внутренняя непротиворечивость имеет гораздо большее значение, чем то, насколько их время близко к реальному. Для таких алгоритмов принято говорить о логических часах (logical clocks).
- В своей классической статье Лампорт (Lamport) показал, что хотя синхронизация часов возможна, она не обязательно должна быть абсолютной.

Отметки времени Лампорта

- Для синхронизации логических часов Лампорт определил отношение под названием «происходит раньше». Выражение a→b читается как «a происходит раньше b» и означает, что все процессы согласны с тем, что первым происходит событие a, а позже — событие b, Отношение «происходит раньше» непосредственно исполняется в двух случаях.
  - Если a и b — события, происходящие в одном и том же процессе, и a происходит раньше, чем b, то отношение a → Ь истинно.
  - Если a — это событие отсылки сообщения одним процессом, а b — событие получения того же .



Три процесса, каждый с собственными часами, которые ходят с разной скоростью (а).

Подстройка часов по алгоритму Лампорта (б)

Алгоритмы выборов

Алгоритмы голосования

- Многие распределенные алгоритмы требуют, чтобы один из процессов был координатором, инициатором или выполнял другую специальную роль.
- Обычно не важно, какой именно процесс выполняет эти специальные действия, главное, чтобы он вообще существовал.
- Роль координатора может выполнять любой процесс.
- Любой процесс может инициировать процедуру выборов.
- Каждый процесс имеет некоторый уникальный номер.
- В общем, алгоритмы голосования пытаются найти процесс с максимальным номером и назначить его координатором. Алгоритмы различаются способами поиска координатора.

Алгоритм забияки

- Когда один из процессов замечает, что координатор больше не отвечает на запросы, он инициирует голосование. Процесс, например P, проводит голосование следующим образом.

  1. P посылает всем процессам с большими, чем у него, номерами сообщение

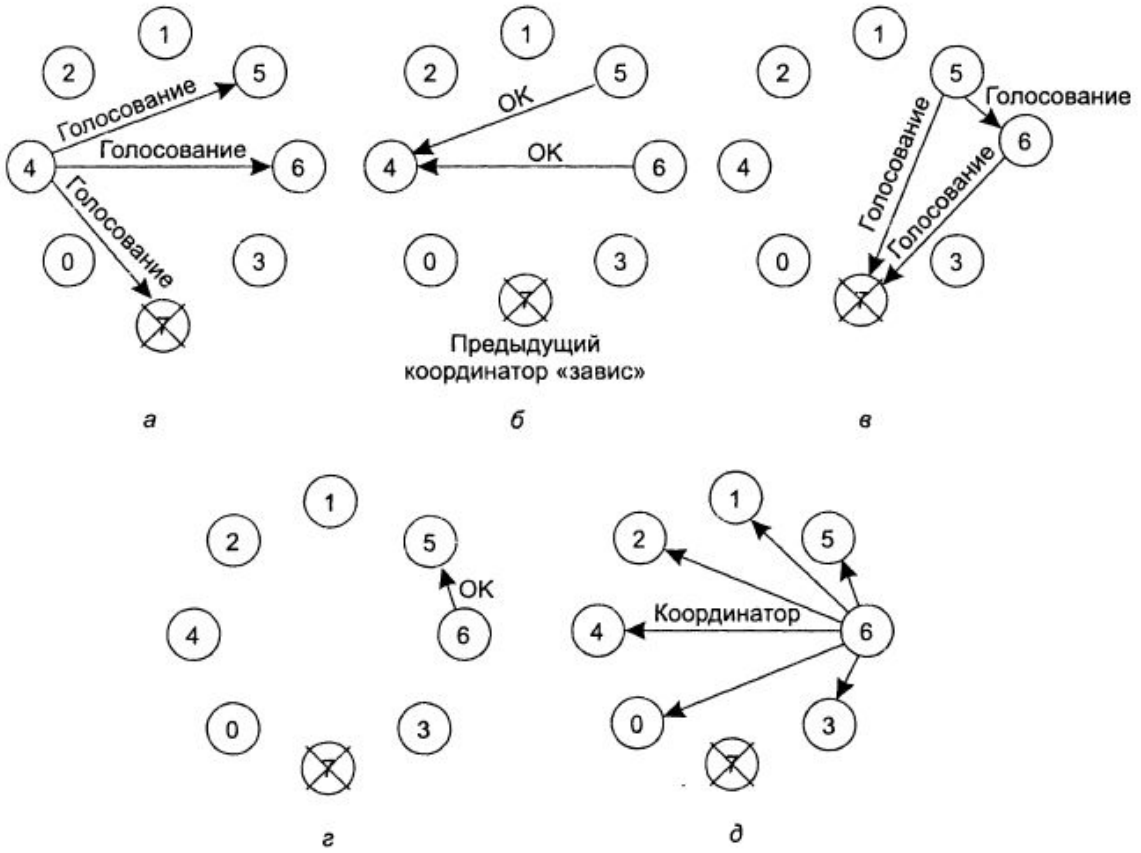
  ГОЛОСОВАНИЕ.

  2. Далее возможно два варианта развития событий:

  - если никто не отвечает, P выигрывает голосование и становится координатором;
  - если один из процессов с большими номерами отвечает, он становится координатором, а работа P на этом заканчивается.

- В любой момент процесс может получить сообщение ГОЛОСОВАНИЕ от одного из своих коллег с меньшим номером. По получении этого сообщения получатель посылает отправителю сообщение ОК, показывая, что он работает и готов стать координатором.
- Затем получатель сам организует голосование. В конце концов, все процессы, кроме одного, отпадут, этот последний и будет новым координатором. Он уведомит о своей победе посылкой всем процессам сообщения, гласящего, что он новый координатор и приступает к работе.

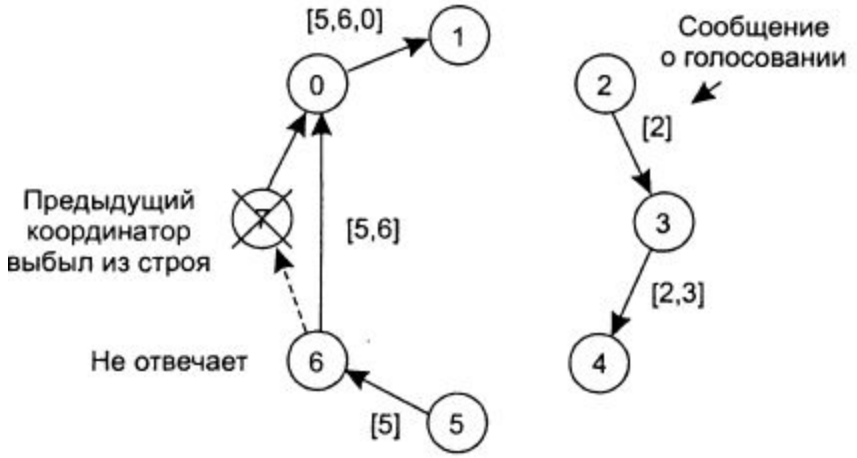
Голосование по алгоритму забияки



- Ранее координатором был процесс 7, но он завис.
- Процесс 4 первым замечает это и посылает сообщение *ГОЛОСОВАНИЕ всем процессам с номерами больше, чем у него, то есть процессам* 5, 6 и 7 (*а*).
- Процессы 5 и 6 отвечают *ОК (б)*
- Оставшиеся процессы, 5 и 6, продолжают голосование (в).
- Каждый посылает сообщения только тем процессам, номера у которых
- процесс 6 сообщает процессу 5, что голосование будет вести он (г).
- В это время 6 понимает, что процесс 7 мертв, а значит, победитель — он сам (д).

Кольцевой алгоритм

- Этот алгоритм голосования основан на использовании логического кольца - процессы физически или логически упорядочены, так что каждый из процессов знает, кто его преемник.



- Когда два процесса, 2 и 5, одновременно обнаруживают, что предыдущий координатор, процесс 7, перестал работать.
- Каждый из них строит сообщение *ГОЛОСОВАНИЕ и запускает это сообщение* в путь по кольцу независимо от другого.
- Оба выбирают 6: [5,6,0,1,2,3,4] [2,3,4,5,6,0,1]

- Когда один из процессов обнаруживает, что координатор не функционирует, он строит сообщение *ГОЛОСОВАНИЕ, содержащее* его номер процесса, и посылает его своему преемнику.
- Если преемник не работает, отправитель пропускает его и переходит к следующему элементу кольца или к следующему, пока не найдет работающий процесс.
- На каждом шаге отправитель добавляет свой номер процесса к списку в сообщении, активно продвигая себя в качестве кандидата в координаторы.
- В конце концов, сообщение вернется к процессу, который начал голосование.
- В этот момент тип сообщения изменяется на *КООРДИНАТОР и вновь* отправляется по кругу, на этот раз с целью сообщить всем процессам, кто стал координатором (элемент списка с максимальным номером) и какие процессы входят в новое кольцо.

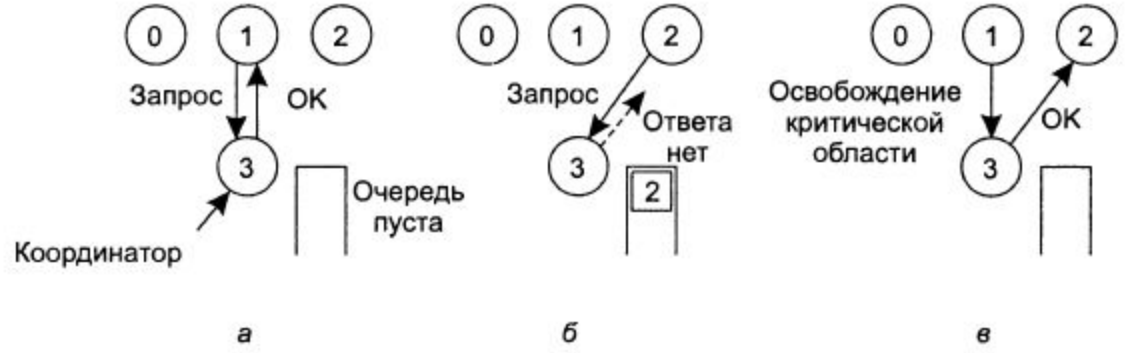
Взаимное исключение

Понятие критической области при работе с ресурсами

- Системы, состоящие из множества процессов, обычно проще всего программировать, используя критические области.
- Когда процесс нуждается в том, чтобы считать или обновить совместно используемые структуры данных, он сначала входит в критическую область, чтобы путем взаимного исключения убедиться, что ни один из процессов не использует одновременно с ним общие структуры данных.
- В однопроцессорных системах критические области защищаются семафорами, мониторами и другими конструкциями подобного рода.

Взаимные исключения в распределенных системах

- Наиболее простой способ организации взаимных исключений в распределенных системах состоит в том, чтобы использовать методы их реализации, принятые в однопроцессорных системах.
- Один из процессов выбирается координатором (например, процесс, запущенный на машине с самым большим сетевым адресом).
- Каждый раз, когда этот процесс собирается войти в критическую область, он посылает координатору сообщение с запросом, в котором уведомляет, в какую критическую область он собирается войти, и запрашивает разрешение на это. (а).



- Другой процесс, 2, запрашивает разрешение на вход в ту же самую область (*б). Координатор знает, что в этой критической* области уже находится другой процесс, и не дает разрешения на вход.
- Когда процесс 1 выходит из критической области, он сообщает об этом координатору, который разрешает доступ процессу 2 (в).

Распределенный алгоритм (1)

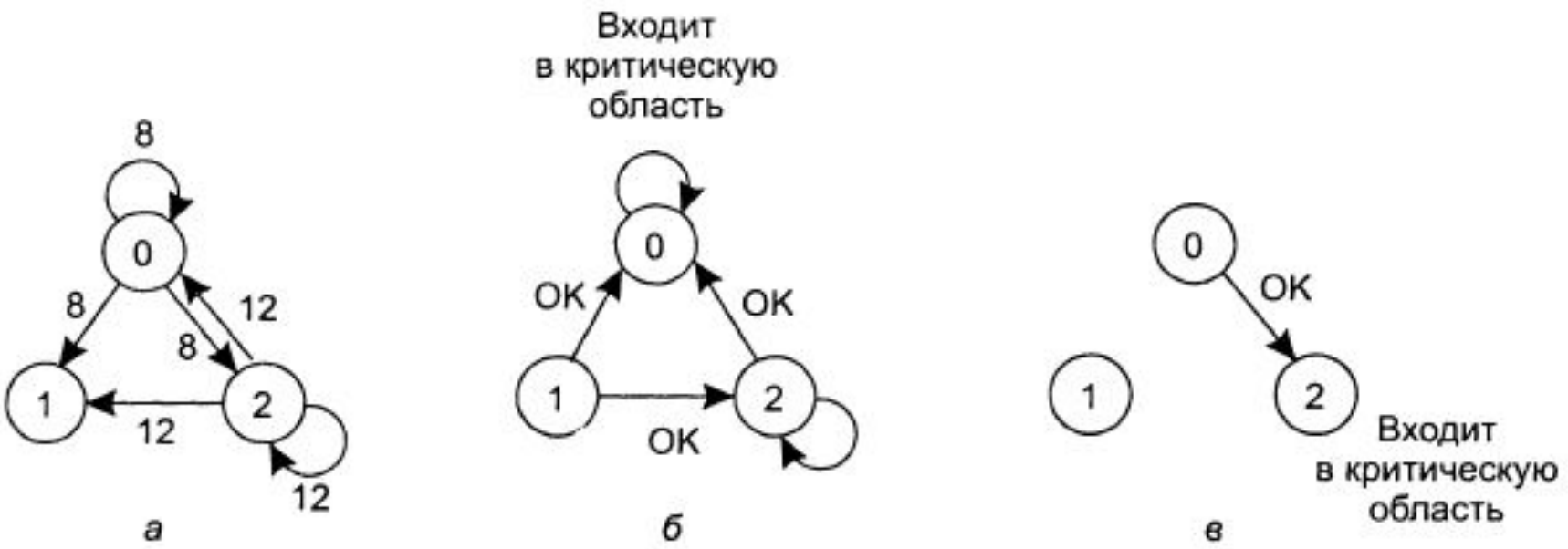
- Рассматриваемый алгоритм требуют наличия полной упорядоченности событий в системе. То есть в любой паре событий, например отправки сообщений, должно быть однозначно известно, какое из них произошло первым. Алгоритм Лампорта, является одним из способов введения подобной упорядоченности и может быть использован для расстановки отметок времени распределенных взаимных исключений.
- Когда процесс собирается войти в критическую область, он создает сообщение, содержащее имя критической области, свой номер и текущее время. Затем он отсылает это сообщение всем процес-сам, концептуально включая самого себя. Посылка сообщения, как предполагается, надежная, то есть на каждое письмо приходит подтверждение в получении. Вместо отдельных сообщений может быть использована доступная надежная групповая связь.
- Когда процесс получает сообщение с запросом от другого процесса, действие, которое оно производит, зависит от его связи с той критической областью, имя которой указано в сообщении.

Распределенный алгоритм (продолжение)

- Можно выделить три варианта:
  - Если получатель не находится в критической области и не собирается туда входить, он отсылает отправителю сообщение *ОК,*
  - *Если получатель находится в критической области, он не отвечает, а помещает* запрос в очередь.
  - Если получатель собирается войти в критическую область, но еще не сделал этого, он сравнивает метку времени пришедшего сообщения с меткой времени сообщения, которое он отослал. Выигрывает минимальное. Если пришедшее сообщение имеет меньший номер, получатель отвечает посылкой сообщения *ОК. Если его собственное сообщение имеет меньшую отметку* времени, получатель ставит приходящие сообщения в очередь, ничего не посылая при этом.
- После посылки сообщения-запроса на доступ в критическую область процесс приостанавливается и ожидает, что кто-нибудь даст ему разрешение на доступ. После того как все разрешения получены, он может войти в критическую область.
- Когда он покидает критическую область, то отсылает сообщения *ОК всем* процессам в их очереди и удаляет все сообщения подобного рода из своей очереди.

Работа алгоритма распределенного исключения

- Представим себе, что два процесса пытаются одновременно войти в одну и ту же критическую область (а).
- Процесс 0 имеет меньшую отметку времени и потому выигрывает (б).
- Когда процесс 0 завершает работу с критической областью, он отправляет сообщение ОК, и теперь процесс 2 может войти в критическую область *(в)*

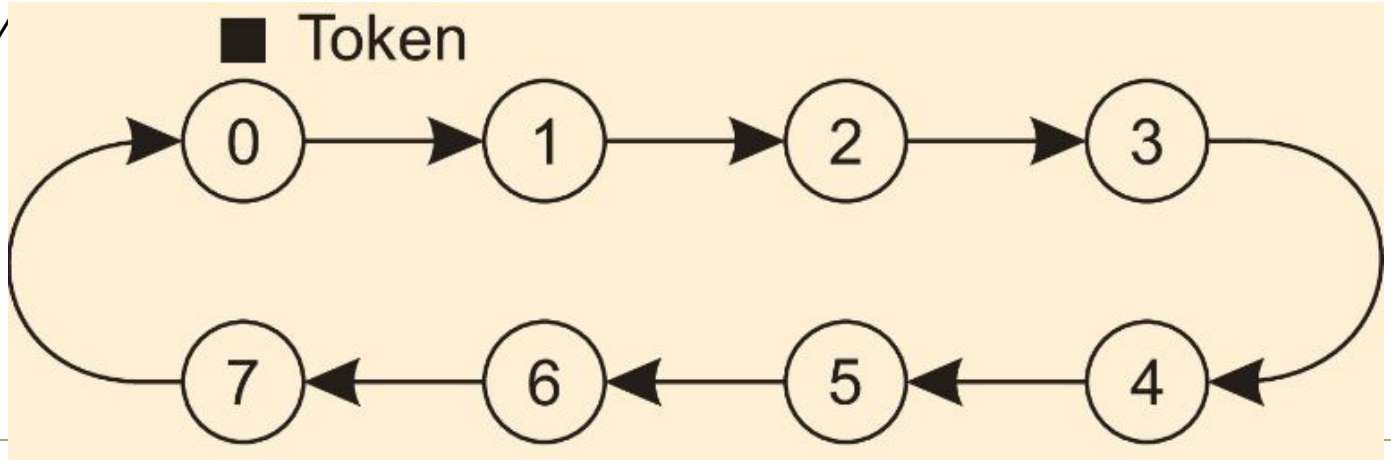


Проблемы распределенного алгоритма

- Если какой-либо из процессов «рухнет», он не сможет ответить на запрос. Это молчание будет воспринято (неправильно) как отказ в доступе и блокирует все последующие попытки всех процессов войти в какую-либо из критических областей.
- Этот алгоритм может быть исправлен так, что когда приходит запрос, его получатель посылает ответ всегда, разрешая или запрещая доступ. Всякий раз, когда запрос или ответ утеряны, отправитель выжидает положенное время и либо получает ответ, либо считает, что получатель находится в нерабочем состоянии. После получения запрещения отправитель ожидает последующего сообщения *ОК*
- Другая проблема этого алгоритма состоит в том, что либо должны использоваться примитивы групповой связи, либо каждый процесс должен поддерживать список группы самостоятельно, обеспечивая внесение процессов в группу, удаление процессов из группы и отслеживание сбоев. Метод наилучшим образом работает, когда группа процессов мала, а членство в группе постоянно и никогда не меняется.
- В распределенном алгоритме все процессы вынуждены участвовать во всех решениях, касающихся входа в критические области. Если один из процессов оказывается неспособным справиться с такой нагрузкой, маловероятно, что возымеет успех попытка их всех сделать то же самое параллельно.
- Алгоритм можно модифицировать так, чтобы разрешить процессу вход в критическую область после того, как он соберет разрешения простого большинства, а не всех остальных процессов.
- Несмотря на все возможные улучшения, этот алгоритм остается более медленным, более сложным, более затратным и менее устойчивым, чем исходный централизованный

Алгоритм маркерного кольца

- Программно создается логическое кольцо, в котором каждому процессу назначается его положение в кольце. При инициализации кольца процесс О получает маркер, или токен (token).
- Маркер циркулирует по кольцу. Он передается от процесса k процессу k+ 1 (это модуль размера кольца) сквозными сообщениями. Когда процесс получает маркер от своего соседа, он проверяет, не нужно ли ему войти в критическую область.
- Если это так, он входит в критическую область, выполняет там всю необходимую работу и покидает область. После выхода он передает маркер дальше.
- Входить в другую критическую область, используя тот же самый маркер, запрещено.
- Если процесс, получив от соседа маркер, не заинтересован во входе в крити



Распределенные транзакции

Распределенные транзакции

- Концепция транзакций тесно связана с концепцией взаимных исключений.
- Алгоритмы взаимного исключения обеспечивают одновременный доступ не более чем одного процесса к совместно используемым ресурсам.
- Транзакции, в общем, также защищают общие ресурсы от одновременного доступа нескольких параллельных процессов.
- Однако транзакции могут и многое другое:
  - Они превращают процессы доступа и модификации множества элементов данных в одну атомарную операцию;
  - Если процесс во время транзакции решает остановиться на полпути и повернуть назад, все данные восстанавливаются с теми значениями и в том состоянии, в котором они были до начала транзакции.

Модель транзакций

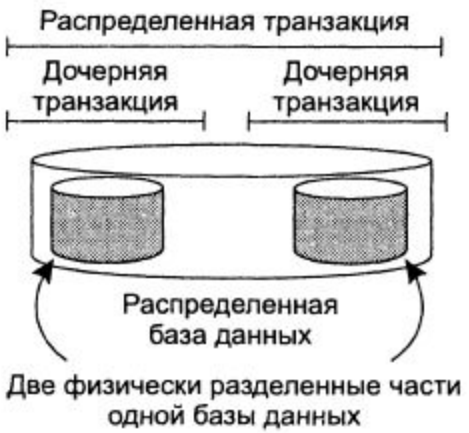
- Свойство транзакций «все или ничего» — это лишь одно из характерных свойств транзакции. Говоря более конкретно, транзакции:
  - атомарны (atomic) — для окружающего мира транзакция неделима;
  - непротиворечивы (consistent) — транзакция не нарушает инвариантов системы;
  - изолированы (isolated) — одновременно происходящие транзакции не влияют друг на друга;
  - долговечны (durable) — после завершения транзакции внесенные ею изменения становятся постоянными.

Примитив	Описание
BEGIN_TRANSACTION	Пометить начало транзакции
END_TRANSACTION	Прекратить транзакцию и попытаться завершить ее
COMMIT	Подтвердить транзакцию
ROOLUP	Отменить (откатить) транзакцию

Классификация транзакций

- Плоская транзакция - серия операций, удовлетворяющая свойствам ACID.
- Плоские транзакции имеют одно ограничение - они не могут даватьчастичного результата в случае завершения или прерывания. Другими словами, сила атомарности плоских транзакций является в то же время и их слабостью.
- Вложенные транзакции: Транзакция верхнего уровня может разделяться на дочерние транзакции, работающие параллельно, на различных машинах, для повышения производительности или упрощения программирования.
- Распределенн… …ноские) работают с да… …м, то такие транзак… *…анзакций (distributed trar…*

а



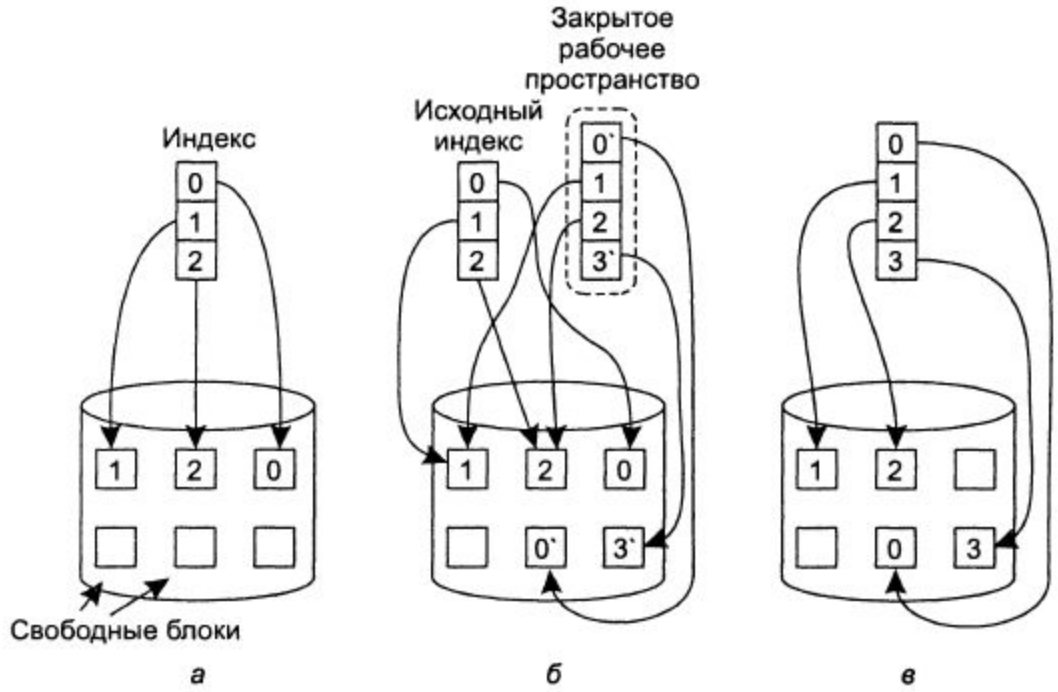
б

Способы реализации транзакций

- Обычно используются два метода:
  - **Закрытое рабочее пространство**.
  - **Журнал с упреждающей записью.**

Закрытое рабочее пространство

- Концептуально, когда процесс начинает транзакцию, он получает закрытое рабочее пространство, содержащее все файлы, к которым он хочет получить доступ.
- Пока транзакция не завершится или не прервется, все операции чтения и записи будут происходить не в файловой системе, а в **закрытом рабочем пространстве**.
- Это утверждение прямо приводит нас к первому методу реализации — созданию д абочего пространства.



- Индекс файла и дисковые блоки для файла из трех блоков (а).
- Ситуация после того, как транзакция модифицировала блок 0 и добавила блок 3 (б).
- Ситуация после подтверждения транзакции *(в)*

Журнал с упреждающей записью

- Согласно этому методу файлы действительно модифицируются там же, где находятся, но перед тем, как какой-либо блок действительно будет изменен, в журнал заносится запись со сведениями о том, какая транзакция вносит изменения, какой файл и блок изменяются, каковы прежние и новые значения.
- Только после успешной записи в журнал изменения вносятся в файл.

Листинг 5.3. Транзакция

```
X = 0:
У = 0:
BEGINJRANSACTION:
X = X + I:
У = У + 2;
X = y * y:
ENDJRANSACTION;
```

Для каждой из трех инструкций тела транзакции до начала ее выполнения создается запись в журнале, которая содержит прежнее и новое значения, разделенные косой чертой:

- содержимое журнала перед выполнением первой инструкции (x=x+1:):

[x=0/1]

- содержимое журнала перед выполнением второй инструкции (y=y+2:):

[x=0/1]
[y=0/2]

- содержимое журнала перед выполнением третьей инструкции (x=y*y;):

[x=0/1]
[y=0/2]
[x=1/4]

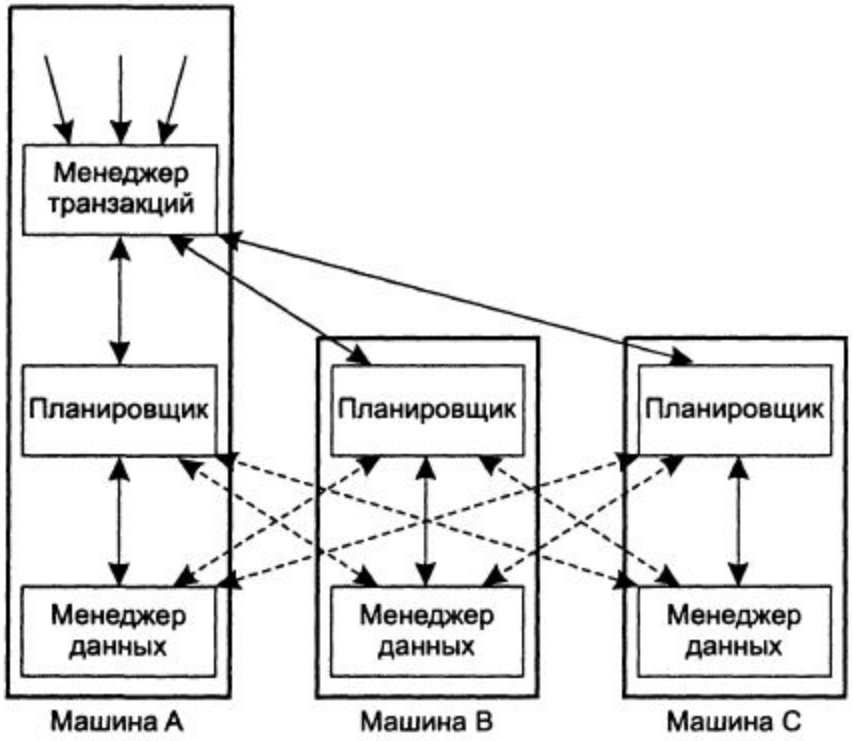
Управление параллельным выполнением транзакций

- Цель управления параллельным выполнением транзакций состоит в том, чтобы позволить нескольким транзакциям выполняться одновременно, но таким образом, чтобы набор обрабатываемых элементов данных (например, файлов или записей базы данных) оставался непротиворечивым.
- Непротиворечивость достигается в результате того, что доступ транзакций к элементам данных организуется в определенном порядке так, чтобы конечный результат был таким же, как и при выполнении всех транзакций последовательно.
- Управление параллельным выполнением лучше всего можно понять в терминах трех менеджеров, организованных по уровням:

На самом верхнем уровне находится *менеджер транзакций.*
На среднем уровне – планировщик.
В самом низу – менеджер данных.

Менеджеры транзакций

- Менеджер транзакций отвечает, прежде всего, за атомарность и долговечность.
- Он обрабатывает примитивы транзакций, преобразуя их в запросы к планировщику.



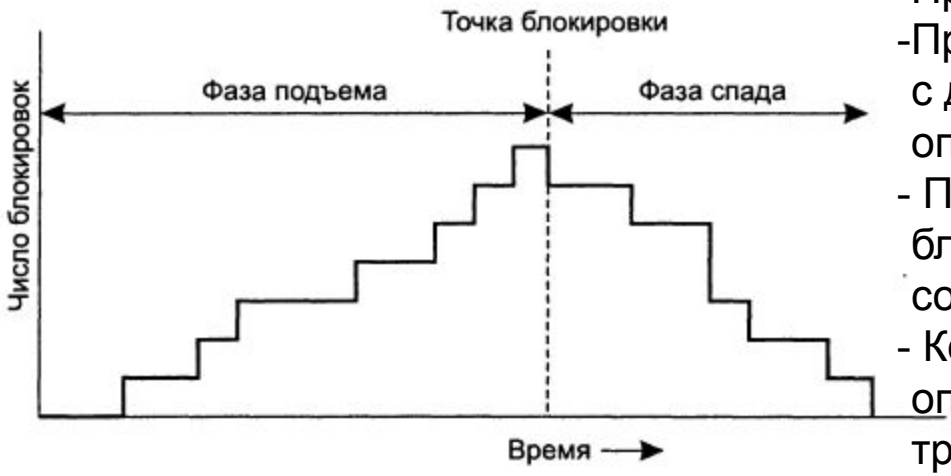
- Каждая машина в этом случае имеет своих планировщика и менеджера данных, которые совместно обеспечивают гарантии непротиворечивости локальных данных.
- Каждая транзакция обрабатывается одним менеджером транзакций. Последний работает с планировщиками отдельных машин.
- В зависимости от алгоритма управления параллельным выполнением транзакций планировщик также может работать с удаленными менеджерами данных.

Изолированность

- Основная задача алгоритмов управления параллельным выполнением — гарантировать возможность одновременного выполнения многочисленных транзакций до тех пор, пока они изолированы друг от друга. Это значит, что итоговый результат их выполнения будет таким же, как если бы эти транзакции выполнялись одна за другой в определенном порядке.

Двухфазная блокировка

- Самый старый и наиболее широко используемый алгоритм управления параллельным выполнением транзакций — это *блокировка (locking).*
- При *двухфазной блокировке (Two-Phase Locking, 2PL),* планировщик сначала, на *фазе подъема {growingphase), устанавливает* все необходимые блокировки, а затем, на *фазе спада {shrinking phase),* снимает их.



При 2PL выполняются три правила:
-Проверка не конфликтует ли эта операция с другими уже заблокированными операциями.
- Планировщик никогда не снимает блокировку с элемента x, если менеджер сообщает, что он выполняет операцию с x.
- Когда планировщик снимает блокировку с операции установленную по требованию транзакции T, он никогда не делает новую блокировку по требованию этой транзакции.

Доказано, что если все транзакции используют двухфазную блокировку, любой план, сформированный путем перекрытия этих транзакций, сериализуем. В этом причина популярности двухфазной блокировки.

Пессимистическое упорядочение по отметкам времени