

Содержание

1 Intro.	2
2 История x86.	2
Немного офф-топ о том, кто использует ассемблер.	3
Продолжаем историю.	3
AMD.	4
3 Займёмся делом.	4
3.1 Регистры.	5
3.2 Адресация.	6
3.3 Инструкции.	6
Присваивания.	6
mov и его друзья.	6
xchg.	6
bswap и movbe.	7
lea.	7
cwd и друзья.	7
Арифметика.	7
Сложение и вычитание.	7
Умножение.	7
Деление.	7
Инкремент и декремент.	8
peg.	8
Логика.	8
хог для обнуления.	8
cmp и test	8
Сдвиги.	8
Работа со стеком.	8
push/pop.	8
pusha(d)/pora(d).	8
Передача управления.	8
Условные переходы.	9
Вызовы функций.	9
Прерывания.	9
por.	9
ud2.	9
Как правильно писать if.	9
Как писать циклы.	10
Как писать switch.	11
4 Немного про оптимизации.	12
Время.	12
Ещё важное.	13
5 FPU.	13
5.1 Регистры.	13
5.2 Инструкции пересылки.	13
FLD, FILD, FBLD.	13
FST(P), FIST(P), FBST(P).	13
FXCH.	13
FCMOV _{cc}	13
5.3 Инструкции арифметики.	13
FADD, FADDP, FIADD.	13

FSUB, FSUBP, FISUB; FSUBR, FSUBRP, FISUBR.	14
FMUL, FMULP, FIMUL.	14
FDIV, FDIVP, FIDIV; FDIVR, FDIVRP, FIDIVR.	14
FPREM(1).	14
FABS.	14
FCHS.	14
FRNDINT.	14
FSQRT.	14
FSIN.	14
FCOS.	14
FSINCOS.	14
FPTAN.	14
FPATAN.	14
FCOMI(P).	14
FINCSTP, FDECSTP.	14
FFREE.	14
FNOP.	14
5.4 Конвенции вызовов (32 bit).	14

1 Intro.

Напоминание: ассемблер — такой язык, который сильно отличается от остальных тем, что полностью зависит от процессора. Ассемблер — набор мнемоник для процессора. А значит ассемблеров столько же, сколько и ISA.

На нашем курсе будет изучаться x86. Помимо него можно встретить ARM (хотя это семейство ISA, они не совместимы друг с другом), MIPS (старое, встречается в роутерах и первых двух PlayStation, из роутеров замещается т.к. мертво), Power (старые маки). А ещё есть новая ISA: RISC-V. Она очень хорошо если захватит мир (т.к. это открытая ISA), но пока нет. Сейчас она очень быстро захватывает контроллеры. Например, контроллеры винчестера. Раньше там были ARM'ы. С ARM сейчас все активно убегают, потому что они отозвали лицензию Huawei, и все остальные боятся, что у них тоже отзовут. Кроме Apple, которые участвовали в создании ARM. Но это точно не всё. Потому что есть видеокарты, которые вполне можно считать специализированными устройствами со своими ISA. В линейке новых видеокарт AMD публикуют документацию к их ассемблеру. И ровно также существует тьма специализированных популярных ISA, но мы их касаться не будем.

При этом важно следующее. ISA и ассемблер сопоставляются друг другу почти один к одному. Одно только: в x86 есть два **синтаксиса** ассемблера: Intel и AT&T+GNU. Первый **ADD RAX, RBX** — прибавление **RBX** к **RAX** и запись в **RAX**. GNU же выглядит как-то так: **ADDL %RBX, %RAX**. То есть у команды уточнена битность (зачем-то), поменяны аргументы местами и стоят проценты. Мы будем использовать первый, а наш ассемблер (как программа) будет `nasm`.

2 История x86.

x86 странноват по историческим причинам. Сначала был 16-битный процессор i8086. К тогда мелкой кампании Intel пришла корпорация IBM и заказала процессоры. i8086 был середнячком того времени. Он уже был обременён совместимостью с 8-битным i8080. Intel позиционировала это как временную меру, но то крутое, что она хотела сделать, так и не состоялось. Что интересно про i8086 — 20-битная адресация. Как это делать с 16-битными регистрами? Сложно сказать.

В итоге i8086 попал в очень популярную линейку компов IBM, где был MS-DOS. Короче, i8086 просто повезло. Потом Intel выпустила i8088 — более дешёвую версию, у которого была 8-битная шина, поэтому он работал медленнее. За i8086 вышли его настоящие последователи: i80186, i80286, i80386, i80486... Всё это обратно совместимо баг в баг. Отсюда и название x86.

Совместимость была лютейшая, и вот промер: команда, которая считает синус, имела баг. При числе x в районе π его синус равен примерно $x - \pi$. Значит надо вставить точную константу π . Но оказалось,

что константа недостаточно точная. Это поправили AMD, пользователи подняли крик, и они вернули старую неточную константу. Впрочем, практически это никого не напрягает, потому что там точность — относительное понятие (потому что совсем точно знать ответ бесполезно, потому что **double** уже скорее всего округлён). Или другой пример: до сих пор современные x86 запускается в начале именно в режиме i8086.

В i80286 добавили защищённый режим, в i80386 — 32-битный режим. В i80486 не скажу, что было, а потом был Pentium (архитектура P5), который стал первым суперскалярным процессором (что, конечно, на ISA не повлияло). Из интересного, Pentium сейчас используется для разных экспериментов. Например, когда они показывали процессор на солнечной батарее, это был именно Pentium. Дальше был Pentium MMX, в котором добавили первое расширение на SIMD — single instruction multiple data. Предназначено для обработки мультимедиа (например, там есть встроенное сложение с насыщением), или любых других больших массивов данных. Отсюда и название, MultiMedia Extension. Параллельно с этим был Pentium Pro (архитектура P6) (out-of-order superscalar, в отличие от Pentium, который in-order superscalar), но он дорогой, а значит непопулярный.

Немного офф-топ о том, кто использует ассемблер. SIMD (и в целом модные расширения, в которые компилятор не очень умеет) — одна из основных причин писать на ассемблере, потому что писать что-то целиком на ассемблере бессмысленно — хороший компилятор вам сильно поможет. Другие причины:

- Сделать что-то, что умеет ОС, и что не умеет ваш язык.
- Системы защиты, вирусы и антивирусы. Код там, кстати, похожий, лол. Более того, они иногда лезут в машинный код, кодогенерацию на лету, более того код может даже полиморфным получаться. (Так Denovo работает.)
- Низкоуровневые оптимизации. Если вы компилятор пишете, например. Если вы пишете код с арифметикой с насыщением. Вопрос: догадается ли компилятор, что вам надо использовать инструкцию? Можно напрямую писать на ассемблере, если вы переписываете какую-нибудь функцию, которая занимает существенное время.
Что интересно, вы обычно пишете несколько ассемблерных вставок под одно и то же. Почему? Потому что ISA разные.
Обычно речь идёт именно про SIMD. Не только потому, что его сложно заметить в высокоуровневом коде, но ещё и потому, что обычно SIMD происходит в числах с пониженной точностью. А значит ваш язык может просто не уметь выражать то, что вы хотите.
- Отладка программ. А вы думали иначе? Если есть исходники, то всё круто. Если нет, со знанием ассемблера вы всё равно можете. В т.ч. когда происходят баги компилятора.

Продолжаем историю. Pentium MMX был первым процессором, на котором можно было смотреть видео 320×240 30 кадров в секунду в MPEG-I без аппаратных ускорителей. После него появился Pentium 2 (который позиционировался как помесь Pentium MMX и Pentium Pro; это правда, но процессор подешевле, конечно, чем Pro) и Pentium 3. Там появился SSE — своеобразный SIMD для вещественных чисел. Потом произошла неудача с Pentium 4 и SSE2 (и SSE3). Если до этого микроархитектура была одинаковой и работающей, то Pentium 4 с новой микроархитектурой дико грелся (по тем временам) и выдавал низкую скорость (из-за извращённой архитектуры) несмотря на адские мегагерцы.

Почему нет 64-битного режима? Потому что во время Pentium 4 Intel хотели выпустить Intel Itanium, свой, новый и с лицензией только у себя. Но он по многим причинам был окрещён Intel Itanic и утонул. В итоге Intel поддержали AMD-шное расширение на 64 бита. Спойлер: оно называется amd64.

Параллельно был Pentium M — развитие Pentium 3 с оптимизацией в энергопотреблении. Это для ноутбуков. Когда стало ясно, что Pentium 4 — это провал, инженеры взяли Pentium M и сделали из него Core 2, в котором добавили расширение SSSE3. Это расширение уже про целые числа, и довольно полезные. В более поздних ревизиях было SSE 4.1. Core 2 был хорош.

Дальше была линейка i7 с микроархитектурой Nehalem (из которых уже потом были урезанные i5),

и там добавили SSE4.2. Дальше серия ix-2xxx и микроархитектура Sandy Bridge, где добавили AVX., потом i7-4xxx (микроархитектура Haswell), где появился AVX2 — расширили регистры SIMD до 256 бит. А ещё там FMA — возможность сделать $d = a * b + c$ с только одним округлением в конце. Она хороша, но считается и за сложение, и за умножение, потому что увеличивает вдвое FLOPS. Потом была серия ix-6xxx (в этот момент начинают расходиться цифры для десктопов и всего остального), так что будем смотреть по микроархитектуре (тут Skylake). Тут было AVX512, но было оно только у серверного Skylake. Обещали, что добавят в десктопы, когда будет новый техпроцесс (10 нм). Спойлер: нового техпроцесса не произошло.

Отсюда и до 10xxx происходит Skylake.

Что дальше? Серия 11xxx (Cypress Cove), которая быстро умерла, в которой они добавили AVX512. Но тут им дали пинка AMD, потому что они начали отгружать ядра подешёвке (и по деньгам, и по энергии). А Intel так не могли, в энергию не лезет. Поэтому в серии 12xxx надо было откуда-то кучу ядер. Поэтому туда добавили слабые ядра. Откуда их взяли? Из Atom — попытки Intel зайти на мобильный рынок (но провалились). Короче теперь в 12xxx (Golden Cove — сильные ядра + Gracemont — слабые ядра) и 13xxx с 14xxx (Raptor Cove + Gracemont) сериях крутые ядра поставляются вместе с пачкой убогих, которые не умеют AVX512. Поэтому Intel решили не поставлять AVX512 туда, где есть слабые ядра. Из интересного, когда 12xxx только выходили, люди научились в BIOS-е выключать убогие ядра и включать AVX512. Больше так нельзя, Intel теперь аппаратно отключают AVX512.

AMD. Ни одна компания не хочет завязываться на другую компанию. Поэтому когда IBM пришли к Intel, они поставили одно условие: они дадут одну лицензию компании AMD. Есть ещё одна лицензия, но она переходила по рукам, выпускала банкоматы под виндой, а сейчас её просто держат в подвале и не используют. Из интересного компания, которая имела третью лицензию, сделала себе аппаратное перепрограммирование команды, которая говорит, кто производитель процессора. Зачем им прикидываться Intel'ом или AMD? Потому что они обнаружили, что винда не грузится. А в винде было написано, что если Intel, делаем одно, если AMD — другое, а иначе нечто нерабочее. При помощи этого ещё тестировали Intel-компилятор, чтобы доказать, что он не использует расширения (даже если может) на процессорах AMD.

i8086 просто клонировали, а дальше, AMD выпускали то же самое (с названиями Am286, Am386, Am486 и Am5x86), но нестандартными расширениями. Потом был K5, K6 (Pentium 2), K6-2 (в который добавили MMX), и K6-3, в котором появилось расширение 3DNow! (теперь можно считать 2 **float** на MMX; Intel её естественно не поддерживали, а вместо этого скрафтили SSE, в котором можно считать 4 **float**'а). Во время Pentium 3 появился Atlon (3DNow! Pro).

Дальше был Atlon XP, в котором наконец-то появился SSE. Дальше неинтересно до тех пор, пока не появился Atlon64 (времена Pentium 4). И он был сильно круче, чем Pentium, несмотря на низкую частоту. Да ещё и в Atlon64 был 64-битный режим (и SSE2). Зачем 64 бита? Потому что адресное пространство очень маленькое. А значит что? Что все (особенно любители серверов) покупают Atlon64. Вместо уходящего в закат и сомнительного Pentium 4 и откровенно отстойного Intel Itanium.

После этого был какой-то непопулярный кал (Phenom и линейка с названиями строительной техники), а сейчас ядро Zen (линейка Ryzen), в котором убрали всякий мусор (типа 3DNow!) и сделали сразу AVX2. Из интересного — Zen 4, который поддерживает AVX512. Ядра поменьше там тоже есть, но не в обычных процессорах, а в мобильных, причём следующего поколения. Ещё есть Zen 4c: «с» — более плотная паковка деталей Zen 4 + порезанная максимальная частота и порезанный L3 вдвое. За счёт этого туда запилили побольше ядер на кристалл. А потом начали мешать обычные и «с», что сильно круче, чем происходящее у Intel. Дальше Zen 5, которых 3 вида: обычные, «с» и мобильные, и мобильные порезаны.

3 Займёмся делом.

Как закончить программу? Одной командой: **ret**. Одна проблема: надо объяснить, что с этого надо начинать. Проще всего жить по правилам C и сказать, что это конец функции **main**. Можно иначе, но так проще и логичнее. И вот программа

```
main:
    ret
```

Разве что под Windows лучше написать `_main` (64бит такого прикола не имеет). Это уже почти работает: надо только сказать, что `main` должно быть видно снаружи:

```
global main
main:
    ret
```

Возможно, потребуется сказать

```
global main
section .text
main:
    ret
```

Если хочется код возврата дать, то так:

```
global main
section .text
main:
    mov eax, 0
    ret
```

Уже хорошо. Теперь возьмём `nasm` и попросим его скомпилировать это. Правда, надо ему указать, под какую систему компилировать. Например, `-f win32` под Windows. У вас появится объектный файл, который можно скормить линковщику (`gcc`, например).

3.1 Регистры.

Первое: x86 — `reg-ten-2` архитектура. И дальше мы будем говорить про регистры, про которые мы что-то можем знать. Т.е. регистры общего назначения. Они вот такие:

- `ax` — аккумулятор.
- `cx` — счётчик.
- `dx` — данные + расширение аккумулятора.
- `bx` — базовый.
- `bp` — немного по-другому базовый.
- `sp` — указатель в стек.
- `si` — source.
- `di` — destination.

При этом начиная с 32 битов эта специализация ослабла, но не исчезла. Всё описанное тут — 16-битное. Причём для обратной совместимости с 8-битным i8080 к регистрам `*x` можно обратиться не полностью, а по половинкам: `*l` — младшая, `*h` — старшая. При этом при изменении частей другая часть не меняется. Когда началось 32-битное расширение, все регистры расширили до 32 бит, написав `e` в начале имени регистра. Теперь, например, `ax` — младшие 16 бит `eax`. Потом добавили 64-битные версии, да ещё и новых регистров добавили, но о них не сейчас.

Кроме этого и совсем системных существуют два интересных: `eflags` (`rflags`) и `ip` (`eip`, `rip`). Что такое флаги? Когда вы делаете действие, помимо нормального результата устанавливают флажки. Они представлены битами регистра `FLAGS`. Например: сложение может переполниться и тогда он

установит CF — carry flag. Есть, например, ZF — её ставят, например, сложение и вычитание, если в результате ноль. Из ещё полезного SF — sign flag — равен старшему биту результата, а значит что он показывает знак результата.

A ip — instruction pointer. Хранит адрес следующей исполняемой инструкции. Ни к флагам, ни к ip нельзя обратиться напрямую, но некоторые команды могут с ними взаимодействовать.

3.2 Адресация.

Когда в качестве аргумента какой-то команды хочется использовать штуку, которая находится по адресу какому-то, вокруг адреса ставятся квадратные скобки. Что можно использовать в качестве адреса? В разных режимах разное. В 16-битном: сумма произвольного подмножества из: базы (регистра `bx` или `bp`), индекса (`di` или `si`) и сдвига — 16-битной константы. В 32-битном: (любой регистр) + (любой регистр, кроме `esp`) * (число из 1, 2, 4, 8) + (32-битная константа). Опять же, любая часть опускается. В некоторых ассемблерах можно менять штуки местами, в большинстве умножение на 1 можно опустить, а ещё некоторые понимают `[eax * 5]`. Как? Это равносильно `[eax + eax * 4]`. Про 64-битную адресацию не расскажем.

3.3 Инструкции.

Присваивания. Никто из инструкций этого параграфа не меняет флаги.

mov и его друзья. Присваивание.

- В нормальном синтаксисе `mov dst, src` присваивает `src` в `dst`. Один любой из аргументов может быть памятью. Например, `mov eax, [ebx]`. Хм-м, сколько данных тут будет присвоено? 4 байта т.к. `eax` имеет 32 бита. Поэтому из памяти, адресом которой является `ebx` читается 4 байта и пишется в `eax`. Каким образом? Берётся `ebx` и ещё следующие 3 байта, и читаются. В x86 раньше в памяти идут младшие байты числа. Это называется little-endian. В других системах может быть строго наоборот, тогда это называется big-endian. В упоротых системах может быть смесь. В сетевых протоколах, к сожалению, выбран big-endian.

Кстати, на последнем месте может быть константа. Но это может привести к неясности. `mov [eax], 123` — это присваивание какой битности? Можно уточнить: `mov byte [eax], 5` или `mov [eax], byte 5`. Принято первое. Получается, что один байт. Можно также написать `word`, `dword`, `qword` для 16, 32 и 64 бит соответственно. Ещё перед памятью может `ptr` стоять, но это не `asm`, а `masm` или её такой кринж.

- Ещё есть `movsx` и `movzx` — копируют меньшее в большее следующим образом: меньшая часть копируется, остаток заполняется нулём для `movzx` либо знаковым битом для `movsx`. Best practices: если нужно присвоить байт в большой регистр, применение этих команд намного лучше чем `xor` с самим собой + `mov` байта. Это потому, что присваивание регистров размера равного битности системы осуществляется быстрее.
- Ещё есть `cmovcc`. Они, имеют чуть меньше вариантов откуда и куда загружать. Например, нельзя на второй аргумент написать константу. А прикол их в том, что они загружают значение при определённом условии. Например, `cmovz` — присваивает, если стоит ZF и `cmovnz` — если не стоит. Помимо таких простых штук есть ещё комбинации, которые позволяют сравнивать числа. Как сравнивать числа? По сути, вычитанием. Если аргументы равны, будет ZF, если первое меньше второго в беззнаковом типе, то это CF. А если в знакомом, то OF. И можно скомбинировать их в, например, `cmovbe` — меньше либо равно (комбинация с `|| z`).

xchg. Обмен без дополнительной памяти. **Атомарный swap.** А значит что? Что делать exchange по памяти невероятно долго.

bswap и movbe. **bswap**: Берёт 32-битный (в 64-битном режиме можно и 64-битный) регистр, меняет endianness на противоположный. **movbe**: **mov**, но с попутным изменением endianness. **movbe** — относительно новая команда, а следовательно может не быть, и код может не оптимизироваться компилятором без указания процессора в коде.

lea. Имеет два аргумента: регистр и память. Она считает адрес памяти, НЕ ОБРАЩАЕТСЯ В ПАМЯТЬ и просто присваивает адрес в регистр. Типичное применение — трёхаргументное сложение, но насколько оно эффективнее — вопрос.

cwd и друзья. **cwd** — берёт старший бит **ax** и заполняет им **dx**. **cdq**, **cdo** делают то же самое, но с парами **eax:edx** и **rax:rdx**.

Ещё есть **cwde**, **cbw** и **cdqe** — соответственно расширяют знаковым битом **al** до **ax**, **ax** до **eax** и **eax** до **rax**. Зачем — поймёте потом.

Арифметика. Важное замечание: в языках мы привыкли что типы имеют знаковость, а операции — нет. Тут ровно наоборот: регистры и память — это просто биты. А арифметика интерпретирует их как число (знаковое или беззнаковое), и совершает операцию.

Сложение и вычитание. **add** и **sub**. Делают соответственно $+=$ и $-=$, в синтаксисе как **mov**. Попутно ставят флаги. Также есть **adc** — складывает два числа и прибавляет к ним CF в качестве единицы — и **sbb** — не могу это объяснить, короче, у вас есть перенос из прошлого разряда, если установлен CF.

Умножение. Беззнаковое — **mul**, знаковое — **imul**. Принимает один аргумент и умножает его на **ax** (или аналог аналогичной битности). Результат имеет битность вдвое больше. Куда присваивается результат — чуть сложнее:

- **ax** = **al** * arg8.
- **dx:ax** = **ax** * arg16.
- **edx:eax** = **eax** * arg32.
- **rdx:rax** = **rax** * arg64.

Что за непонятное двоеточие? Возьмём два регистра, склеим из них одно число побольше, где первый регистр — старшая часть.

А ещё есть **imul reg, reg** — делает $*$, отбрасывает старшую часть. Почему нет такого же для беззнаковых? А потому что верно следующее утверждение: младшая половина результата получается одинаковой для знакового и беззнакового умножения.

А ещё есть **imul reg, reg, const** — умножает второе на третье, присваивает в первое. Опять же, знаковость на самом деле не важна.

Деление. Есть **div** и **idiv**, и в их классической форме они похожи на правила для умножения. Например, он берёт пару **edx:eax** и делит на аргумент (регистр или память). Частное записывается в **eax**. Что более интересно, он также считает остаток и присваивает в **edx**. Для 16 битов и 64 битов понятно, для 8 — остаток присваивается в **ah**.

Деление на ноль? Это определено как исключительная ситуация, система останавливает выполнение, переходит в ядро к обработчику ОС и начинает делать какой-то другой код. Но это ладно. Мы делим 64-битное число на 32-битное. Оно легко может не влезть в 32-битное. И это тоже считается делением на ноль.

Что делать, если мы хотим делить друг на друга числа одинакового размера? Ну, если числа беззнаковые, то старшую часть надо обнулить. Если знаковые, то заполнить знаковым битом, и тут мы вспоминаем **cwd** и друзей.

Инкремент и декремент. `inc` и `dec`. Делают почти всё как надо, но не меняют флаг переноса. Если вы используете этот факт, они сильно замедляются. Что в них хорошего — занимают меньше памяти, а чем меньше памяти, тем больше команд влезет в кэш.

neg. Меняет знак числа на противоположный.

Логика. `and`, `or`, `xor`. Побитовые логические операции. И `not` — побитовое отрицание.

xor для обнуления. Традиционно для обнуления регистра принято использовать `xor` его самого с собой. Эта инструкция занимает меньше памяти, и более того распознаётся процессорами как некоторая идиома, и если вы сделали, например `xor eax, eax`, то дальше если вы используете `eax`, то его использование не приостанавливает Superscalar.

cmp и test . `cmp` — делает то же самое, что `sub`, но не пишет результат, пишет только флаги. `test` — то же самое, но `and`. Первое используется для сравнений на больше или меньше, `and` — для установленных битов и сравнения с нулём т.к. `cmp` чего-то и нуля занимает много места (т.к. надо записать 32-битную константу 0).

Сдвиги. `shr`, `shl`, `sar`, `sal`.

Это всё логические или арифметические сдвиги. Сдвиги принимают два аргумента: что двигать и насколько. Сдвиг вправо двигает биты на несколько позиций в сторону меньших битов. Самый старший сдвинутый бит перемещается в CF. Логический сдвиг заполняет освободившиеся позиции нулями (а значит полностью аналогичен делению на степень двойки). Арифметический сдвиг заполняет их знаковым битом, что почти аналогично делению. Разница в том, что арифметический сдвиг округляет к $-\infty$. Сдвиги влево одинаковы и сдвигают в другую сторону. В CF пишется младший сдвинутый бит. Тут есть проблема с аргументами. Первым аргументом может являться только регистр или память, а вторым — константа или `cx`. Причём из него берутся только младшие биты в необходимом для этого количестве. Например, при сдвиге 32-битных чисел берутся только 5 бит.

А ещё есть сдвиги двойной точности: `SHLD` и `SHRD`. Принимают на вход два регистра и число, двигает первый регистр в нужную сторону, но пустые места заполняются частью второго аргумента. То есть для сдвига вправо пустые места заполняются младшими битами второго регистра. Для сдвига влево — заполняются старшими битами второго регистра. CF заполняется так же, как и для обычных сдвигов. И наконец `rol` и `ror` — циклический сдвиг битов в одну из сторон. Последний выдвинутый бит всё ещё отправляется во флаг переноса. Ещё в ту же степь есть `rcl` и `rcr`. Возьмём параметр, приклеим к нему CF, сделаем циклический, сука, сдвиг полученного числа, битности на 1 больше, чем аргумент.

Работа со стеком.

push/pop. Пушить на стек можно регистры, память, константы. Это по сути сдвиг `esp` выше по стеку на размер аргумента (то есть вычитание) + запись туда. `pop` принимает один аргумент (регистр или память), куда записать то, что сняли со стека. Если записывать никуда не надо, то что? Ну, можно просто прибавить значение к `esp`.

Всё то не меняет регистры.

Совет: очень хорошо бы, если бы все штуки на стеке были выровнены кратно режиму программы.

pusha(d)/popa(d). Здесь и далее круглые скобки означают, что это дополнительный суффикс. Т.е. тут две пары команд: `pusha`, `pushad` и `popa`, `popad`.

`pusha` пишет на стек все 16-битные регистры. `popa` читает все регистры и восстанавливает все, кроме `sp`. С суффиксом `d` делает то же самое с 32-битными регистрами.

Передача управления. Самый простой — `jmp` — переход по метке, регистру или адресу в памяти.

Условные переходы. `jc`, где `cc` — то, что мы видели выше в разделе с `cmov`. Применимо только к меткам.

Вызовы функций. `call` и `ret`. `call` — принимает на вход метку, регистр или адрес, записывает на стек instruction pointer и делает `jmp` по своему параметру.

`ret` — снимает со стека значение и переходит по нему.

`ret n` — снимает со стека адрес, увеличивает `esp` на `n`, переходит по тому, что в первый раз снял.

Прерывания. `int` — переходит на обработчик прерываний. Предназначено для взаимодействия с железом. `int 80h`, например — позвать ядро (на Linux). Ещё есть парная команда — `iret` т.к. прерывание сохраняет на стек побольше, чем функция (а именно адрес возврата, флаги, иногда стек).

А ещё есть `int3` — отладочное прерывание. Если у вас нет отладчика, вы умираете, если есть, то он её перехватит и начнёт делать что-то своё.

нор. Делает ничего. Формально эквивалентен `xchg eax, eax`. А ещё есть многобайтовые версии `nor`. Применяется, чтобы начало короткого цикла выравнять по границе кэш-линии.

ud2. Команда, которой не существует. Исполняя её, гарантированно упадёт с ошибкой «команда некорректна».

Как правильно писать if. Вот у нас в `if` есть что-то типа `eax > 5`. Чтобы сформулировать, как писать условие на ассемблере, надо сначала понять, как мы сравниваем: знаково или беззнаково. если это сравнение беззнаковое, нам надо использовать прыжки с `a` и `b`. Иначе — `l` и `r`. При этом если в `if` несколько условий, то написание `if`'а на ассемблере отличается в зависимости от того, как связаны условия.

Например:

```
if (eax > 5 && ebx) {
    // then-branch
}
```

трансформируется в

```
cmp eax, 5
jbe .lend
test ebx ebx ; так традиционно сравнивать с нулём само число
jnz .lend
then-branch
```

`.lend:`

`\end{minted}`

Заметим, что тут хорошо работает правило из C про ленивое исполнение. Это важно. Алгоритм может зав

Другой пример:

`\begin{minted}{c}`

```
if (eax > 5 || ebx) {
    // then-branch
}
```

трансформируется в

```
cmp eax, 5
ja .lthen
test ebx ebx
jnz .lend
.lthen
then-branch
.lend:
```

Что делать, если есть else? Да всё просто, переходы на `.lend` заменить на переходы в `else`, а после конца `if` сделать безусловный переход на метку после окончания `if`'а.

Вы не хотите разрывать проверку от условного перехода. Например, потому, что современные процессоры умеют в *macro-fusion*. Обычно одна внешняя команда превращается в одну или несколько микроопераций, в зависимости от наличия железа для выполнения команды как одной. Например, `add eax, [ecx]` свободно может превратиться в две команды: чтение памяти и сложение. Из интересного, эти действия могут происходить в разное время. Так вот *macro-fusion* — это фишка, которая позволяет превращать несколько команд в одну. В x86 такое типично может быть для сравнения + перехода.

Мы уже говорили про то, что порядок условных переходов иногда критичен для алгоритма. Но если нет, этим можно пользоваться. Если есть условный переход, который легко предсказывается, и который позволяет обойти остальные, то его имеет смысл вставлять в начало, если можно.

И ещё *best practise*. Пусть мы хотим сравнить `if (x >= 3 && x < 7)`, где `x` беззнаковый. Это отлично можно сделать за один переход.

```
mov ecx, x
sub ecx, 3
cmp ecx, 4
jae ecx .lend
; then-branch
.lend
```

Но тут вопрос в том, насколько хорошо это предсказывается. Если значения расположены случайно, один переход лучше, чем два. Более того, такой же прикол в C работает.

Что ещё интереснее, то же работает для знаковых типов. На ассемблерах буквально, а в C придётся написать пару букв: `if (x - 3u < 4u)`.

Как писать циклы.

- Начнём с цикла `do { /* X */ } while (eax < 5u);`. Как выглядит на ассемблере? Ну, очевидно:

```
.loop:
; X
cmp eax, 5
jb .loop
```

- Теперь `while (eax < 5u) /* X */`. Окей:

```
cmp eax, 5
jnb .lend
.loop:
; X
cmp eax, 5
jb .loop
.lend
```

То есть компилятор превращает `while` в `do-while` с условием в начале. Более того, если компилятор умеет доказывать, что цикл `while` исполняется хотя бы раз, он может это убрать.

- И, наконец, цикл `for`. `for (eax = 0; eax < ebx; eax++) { /* X */ }`. От `while` он отличается ничем по сути, и его можно переписать так:

```
eax = 0;
if (eax < ebx) {
do {
// X
```

```

        eax++;
    } while (eax < ebx);
}

```

Конкретно в данном случае if можно убрать.

Но это ладно. Немногие знают, что лучше работают циклы в обратном порядке (от числа до нуля). Почему? Потому что `dec` отлично проставляет ZF (но не CF, это важно). Более того, `maskfusion` отлично работает для комбинаций `dec + jz` и `dec + jnz` (и прочих условных переходов, если они не используют CF). И также работает сложение и вычитание с переходами. Более того, компиляторы умеют переворачивать циклы:

```

        mov eax, ebx
.L1:
    X
    dec eax
    jnz .L1

```

Но эта штука идёт от 5 до 1, а не от 4 до 0. Надо починить. Ну, всё просто

```

        lea eax, [ebx - 1]
.L1:
    X
    sub eax, 1
    jnc .L1

```

Это правильный вариант. Но есть ещё более крутой, если `eax` не очень большой.

```

        lea eax, [ebx - 1]
.L1:
    X
    dec eax
    jns .L1

```

Но это ладно. Что если мы хотим двигаться вперёд? Можно двигаться вперёд к нулю. Типа такого:

```

for (eax = -ebx; eax < 0; eax++) { /* X */ }

```

Тут надо пошаманить с X, но если мы там к массиву хотим обратиться, то всё вообще круто. Например, `X := ecx[eax]++`, где `ecx` — массив `dword`'ов.. Тогда делаем так:

```

        lea ecx, [ecx + ebx * 4]
        mov eax, ebx
        neg eax
.L1:
    inc dword [ecx + eax * 4]
    inc eax
    jnz .L1

```

Важное примечание: везде вышенаписанный код опирается на то, что хотя бы раз условие цикла выполнится. Иначе сверху надо бахнуть ещё одно сравнение.

Как писать switch. Для начала: аргументом `switch` может быть только целое число. Веткой `case` — только константа времени компиляции. В честь чего? В честь того, что только так это можно скомпилировать лучше, чем в кучу `if`'ов. Возьмём такой `switch`:

```

switch (eax) {
case 1:
case 2:
    // X
    break;
case 3:
    // Y
case 5:
    // Z
}

```

Так вот, это отлично компилируется в такое:

```

section .text
    cmp eax, 5
    ja .lend
    jmp [table + eax * 4]
.L1:
    ; X
    jmp .lend
.L2:
    ; Y
.L3:
    ; Z
.lend:

section .rdata
.table: dd .lend, .L1, .L1, .L2, .lend, .L3

```

Короче, тут безусловный переход по табличке. Если значения разнородные (типа 1, 2, 4 и 100000), компилятор их распилит. А если они совсем случайные, то будет куча if'ов. Возможно, с двоичным поиском.

4 Немного про оптимизации.

Самое главное, определить проблему.

Время. За время отвечает команда `rdtsc` и её вариации. Например, `rdtscp`. Обе пишут в `edx:eax` время с момента reset в тактах. Вторая портит `ecx`, но её менее вольно можно переставлять в Superscalar'е. При этом кратность может быть больше одного такта. А ещё команда достаточно тяжёлая (около 100 тактов). Вообще зная тактовую частоту, можно пересчитать в секунды. А после того, как процессоры научились разгоняться, незадолго после этого таймер `rdtsc` отвязали от таймера процессора. А частота этого таймера примерно равна базовой частоте процессора. И да. Точность любых других способов очень мала. Тут точность — сотни тактов. В том, что даёт операционная система, точность — миллисекунды.

Как с этим жить? Да на самом деле не очень хорошо. Если Вы измеряете скорость кода и всё, то ладно. Но если вам нужно понять, почему, надо ставить профилировщик. Причём лучше всего, чтобы профилировщик был от того производителя, что и процессор. Правда, у AMD профайлер не очень. Но зато там есть норм такое API доступа ко всяким встроенным счётчикам, которые могут помочь. А ещё можно стабилизировать частоту процессора, но довольно сложно. На Windows такое может `aida64` (как профилировщик она не очень, но там можно попросить выключить турбо). На Linux нужно как-то добраться до MSR и найти, какой флаг отвечает за турбо. Правда, стабилизация частоты, вероятно, сильно снизит скорость работы. Зато `rdtsc` будет более стабильным.

Ещё важное. Чтобы что-то оптимизировать, надо ещё понимать, насколько быстро работают команды. Например, это можно сделать [тут](#). Тут надо знать, что есть две скорости: latency и throughput. Первая — количество тактов от начала вычисления до получения результата. Вторая — сколько команд в единицу времени может исполнять процессор. Концептуально, первое о последовательности зависимых команд, второе — об одновременном количестве исполняемых команд.

5 FPU.

Оно же имеет название x87. Это модуль работы с плавающей точкой. Изначально это был отдельный набор команд и отдельные микросхемы, которые назывались так же, как и процессор, но с семёркой на конце. Такая топология, что FPU вычисляется на сопроцессоре, повлияла на сущность этих команд.

5.1 Регистры.

Служебные мы тут обсуждать не будем. Только общего назначения. Они называются [R0](#), [R1](#) и т.д. до [R7](#), но в таком виде их не применяют. А всё потому, что эти регистры организованы в виде стека. [R0](#) снизу, [R7](#) сверху. А указатель стека находится в определённом месте и называется [ST0](#). Если мы делаем [push](#), то [ST0](#) сдвигается вниз.

И это всё ладно. Но самое плохое, что ячейки помечены как занятые или свободные. И попытка прочитать не занятую или записать в занятую производит NaN.

Все регистры — 80-битные. И в формате extended-precision они хранят число. При этом FPU умеют автоматически конвертировать типы данных. Они поддерживают целые числа **со знаком** (16, 32 и 64 бита), плавающую точку в форматах single, double и extended (кажется, обозначение `tbyte`) и BCD80. BCD80 — это про десятичные цифры в байтиках. Каждые 4 бита хранят десятичную цифру, и байтов 10.

5.2 Инструкции пересылки.

FLD, FILD, FBLD. Загрузка на стек. Имеют один аргумент (адрес в памяти). [FLD](#) ещё умеет в `ST(i)`. А отличаются они тем, что [FLD](#) оперирует числами с плавающей точкой, [FILD](#) — целыми числами, [FBLD](#) — BCD80.

Плюс есть [FLDZ](#), [FLD1](#) и [FLDPI](#), которые загружают константы (0, 1 и π соответственно). Есть ещё несколько, но это можно загрузить.

FST(P), FIST(P), FBST(P). Если без P — взять значение из [ST0](#) и поставить в память или в указанное место на стеке. Если с P, то записать и после этого освободить ячейку и увеличить указатель стека (если делать версию с P и в качестве аргумента дать стек, запись будет по старым адресам).

FXCH. Меняет местами вершину стека и ещё одно значение стека. Она, как и [mov](#), выполняется за 0 тактов.

FCMOV_{cc}. Условная загрузка. Использует флаги процессора.

5.3 Инструкции арифметики.

Всегда: один аргумент — вершина стека. Другой — стек или память. Если команда обращается к памяти, то память — второй аргумент. При этом вершина стека может быть любым аргументом, но по умолчанию он первый.

FADD, FADDP, FIADD. [FADD](#) — складывает и присваивает в первый аргумент. [FADDP](#) — складывает и делает pop со стека. [FIADD](#) — складывает [ST0](#) и целочисленную константу из памяти.

FSUB, FSUBP, FISUB; FSUBR, FSUBRP, FISUBR. Обычное вычитание берёт первый аргумент, вычитает второй и присваивает в первый. Обратное вычитание берёт второй аргумент, вычитает первый и присваивает в первый.

FMUL, FMULP, FIMUL. Понятно.

FDIV, FDIVP, FIDIV; FDIVR, FDIVRP, FIDIVR. Тоже понятно.

FPREM(1). Вычитает **ST1** из **ST0** не более чем 64 раза до тех пор, пока модуль разности не будет меньше модуля **ST1**. Версия с единичкой отличается округлением.

FABS. Присваивает в **ST0** его модуль.

FCHS. Меняет знак **ST0**.

FRNDINT. Округляет к ближайшему чётному.

FSQRT. Берёт квадратный корень.

FSIN. Берёт синус.

FCOS. Берёт косинус.

FSINCOS. Берёт синус и косинус (берёт значение со стека, кладёт, кажется, сначала синус, потом косинус).

FPTAN. Берёт тангенс.

FPATAN. Считает **atan2**. То есть арктангенс **ST1** / **ST0**. Снимает со стека обоих.

FCOMI(P). Сравнивает **ST0** с **ST(i)** и ставит флаги **ZF**, **PF**, **CF**. При этом делает он сравнение как будто беззнаковое. Вариация с **P** снимает значение со стека.

FINCSTP, FDECSTP. Двигает указатель стека на единичку. Не меняет занятость.

FFREE. Пометить **ST(i)** как свободную ячейку.

FNOP. Делает ничего. Но не совсем. Он умеет бросить возникшее, но ещё не брошенное исключение. Да, IEEE-754 умеет в исключения. И железо это умеет, но надо включить это через служебные регистры. А вообще если исключение происходит, происходит системное прерывание (и гипотетически его можно обработать).

5.4 Конвенции вызовов (32 bit).

Аргументы **single** и **double** лежат на стеке как число соответствующей битности. С **extended** сложнее. Возвращаемое значение лежит на вершине стека сопроцессора. При вызове функции стек сопроцессора должен быть чист, при возврате, если возвращаемого значения нет, то стек должен быть чист, если есть, то чист кроме **ST0**, где лежит результат.