Лекция 5. Масштабирование полевых транзисторов в интегральных схемах

Как уже упоминалось, современные цифровые микросхемы основаны на КМОП–технологии, которая экономична и использует комплементарные пары *n*–МОП и *p*–МОП транзисторов. Огромным преимуществом этой технологии является также возможность масштабирования (англ. *scaling*), т. е. пропорционального изменения всех размеров прибора без ухудшения его характеристик.

С того времени, когда был создан первый процессор, в течение нескольких десятилетий размеры всех его элементов и расстояния между ними постоянно уменьшались, а архитектура и структура самого МОП–транзистора при этом практически не менялись.

Причины такой тенденции достаточно очевидны, так как увеличение плотности упаковки элементов позволяет:

1. уменьшить вес и габариты аппаратуры, что особенно важно для устройств, содержащих сотни миллионов элементов;
2. разместить на той же площади большее количество активных элементов и расширить функциональные возможности устройства;
3. уменьшить ослабление сигнала при прохождении его через СБИС, что позволяет снизить рабочие напряжения и токи, уменьшить потребляемую мощность, облегчить условия теплоотвода и продлить время работы устройств с автономными источниками питания;
4. повысить быстродействие за счет уменьшения времени переноса носителей через активные области и межсоединения;
5. увеличить число транзисторов на пластине, обрабатываемых в одном технологическом цикле и, как следствие, снизить стоимость каждого из них, и уменьшить разброс их параметров;
6. увеличить процент выхода годных чипов, т. к. чем больше площадь чипа, тем больше вероятность, что он попадет на участок пластины, содержащий неустранимый дефект; вероятность выхода годных можно оценить по такой формуле: $Y = \frac{1}{\left[1 + S \cdot \left(N_{\text{пл}} + N_{\text{ш}}\right)\right]^n}$, где S — общая площадь поверхности чипа, n — число литографических циклов, а $N_{\text{пл}}$ и $N_{\text{ш}}$ — плотность дефектов (в расчете на единицу площади) в пластине и фотошаблоне.

Рис. 3.1 показывает, как за последние десятилетия изменялись линейные размеры и быстродействие КМОП–транзисторов и ИС на примере продукции корпорации Intel. Приведенные на рисунке значения *проектных норм* — это линейное разрешение литографии, которая использована в данном технологическом цикле (англ. *Technology Node*).

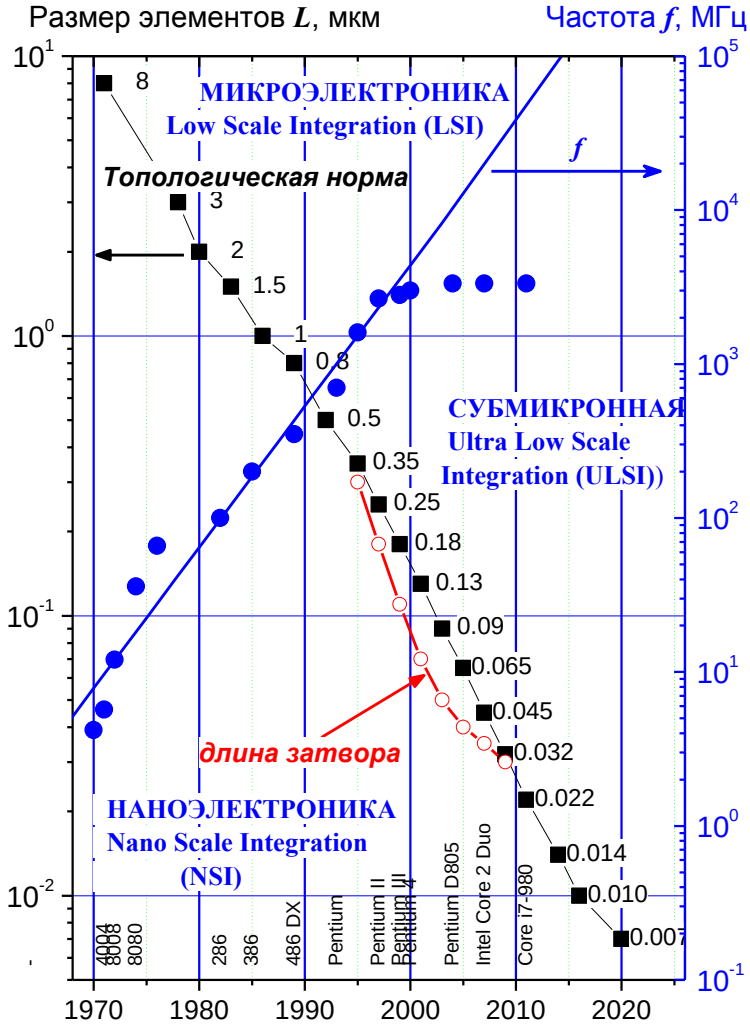

Рис. 3.1. Изменение размеров элементов ИС и рабочей частоты по годам (данные корпорации Intel).

Обычно это значение близко к длине затвора L_G и, разумеется, много меньше реальных размеров самого МОП–транзистора, которые и задают в ИС плотность упаковки.

Эти размеры можно оценить по другому параметру, который в нашей литературе называют «шаг затвора» (англ. *Contacted Gate Pitch*). Связь этого параметра с длиной затвора (и проектной нормой) иллюстрируется на рис. 3.2 по данным той же корпорации Intel. Из рисунка видно, что он в четыре с лишним раза больше литографической нормы.

Проектная норма — это минимальное расстояние между параллельными линиями одинаковой толщины с расстоянием между ними, равным этой толщине, которое еще можно различить на литографическом изображении в рассматриваемой технологии.

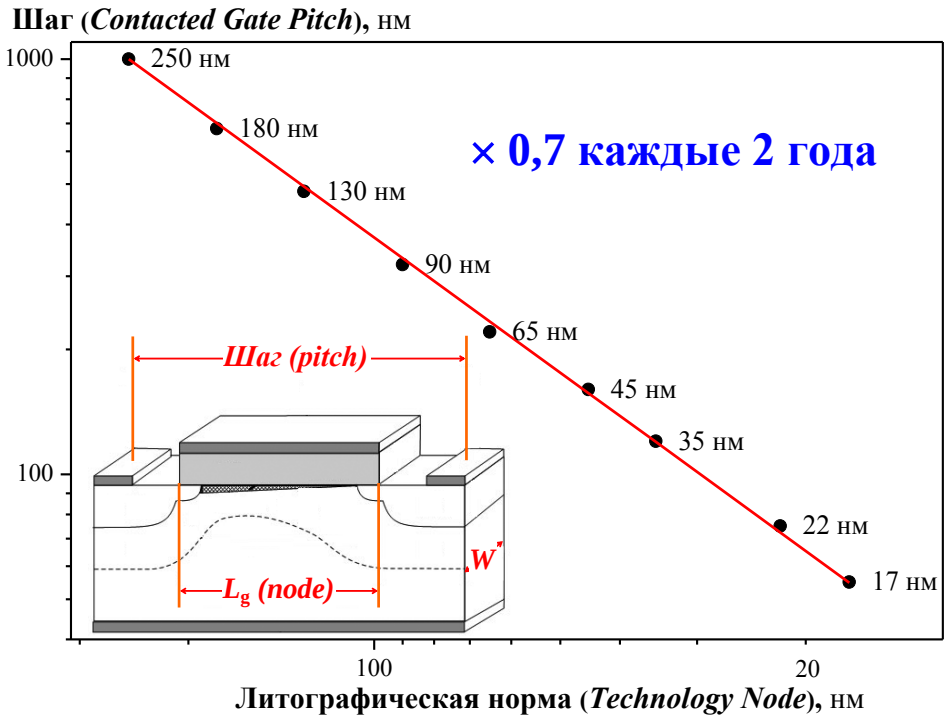

Рис. 3.2. Зависимость шага затвора от проектной нормы фотолитографии (данные корпорации Intel)

3.1. Основные принципы масштабирования (закон Деннарда)

Классическая схема масштабирования предусматривает, что при уменьшении всех размеров элементов микросхемы должно оставаться неизменным электрическое поле *F*. Везде в дальнейшем для обозначения напряженности электрического поля используется латинская буква *F* (от англ. *field*), а не привычная *E*, которой в данном тексте обозначается энергия электронов.

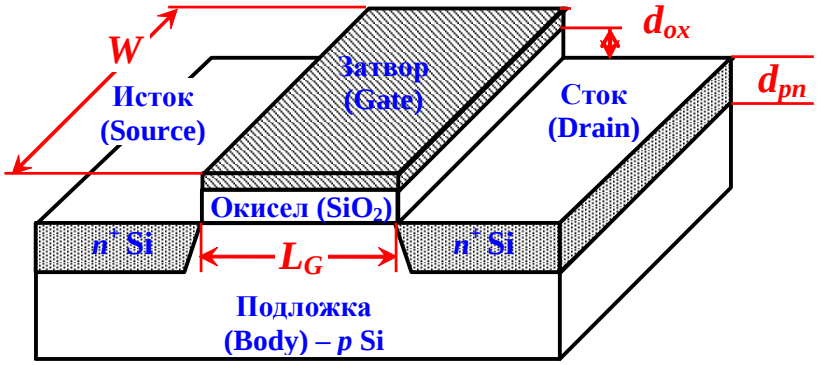

Рис.3.3. Характерные размеры МОП-транзистора, изменяющиеся при масштабировании

Рассмотрим, как должны измениться характеристики МОП–транзистора при масштабировании (*scaling*), — т. е. при уменьшении в *k* раз — всех его геометрических размеров: длины L_G и ширины W_G

затвора, толщины подзатворного окисла d_{ox}, глубины залегания *p-n* перехода d_{pn} (рис. 3.3), а также не показанных на этом рисунке размеров межсоединений.

Впервые теорию масштабирования разработал один из ведущих сотрудников корпорации IBM Роберт Деннард (*R. H. Dennard*), который в 1968 году изобрел полупроводниковую динамическую память с произвольным доступом (англ. *Dynamic Random Access Memory*, DRAM). Эта память во много раз превосходила по емкости и габаритам использовавшуюся тогда ферритовую память и была в 100–1000 раз дешевле. После этого изобретения Деннард возглавил в IBM работу по совершенствованию такой памяти и вместе с коллегами провел многочисленные активные исследования, которые показали, что при уменьшении линейных размеров МОП–транзистора и пропорциональном уменьшении подаваемого на затвор напряжения переключающие свойства транзистора сохраняются, а скорость переключения повышается. Из этого следовало, что для повышения производительности ИС надо увеличивать плотность упаковки и рабочую частоту, а энергопотребление снижать. Такое предсказание не только объясняло рассмотренный ранее закон Мура, но и расширяло его, поскольку в самом законе говорилось только о периодическом уменьшении размеров всех элементов и снижении их стоимости, а производительность процессора вообще не обсуждалась.

Анализ полученных результатов позволил Деннарду и его сотрудникам опубликовать в 1974 году статью, которая вскоре заслужила имя собственное — ее стали называть *Scaling Paper* (статья о масштабировании). Эта статья определила будущее технологии производства цифровых микросхем на несколько десятилетий. В 2007 году вышел специальный номер журнала IEEE Solid-State Circuits Society News под общим заголовком «Влияние и последствия теории масштабирования Деннарда». В номере было опубликовано несколько статей с оценкой роли теории масштабирования и ее автора в истории индустрии информационных технологий, и в оригинале была

R.H. Dennard et al. «Design of ion-implanted MOSFET's with very small physical dimensions» (Проектирование ионно–имплантированных МОП–транзисторов с очень малыми физическими размерами). The IEEE Journal of Solid-State Circuits, v. 9, N 5, 1974, p. 256–268

воспроизведена сама легендарная статья. Сейчас уже не вызывает сомнений, что наряду с законом Мура, указывающим направление развития технологии, необходимо говорить и о законе масштабирования Деннарда, объясняющим, каким именно образом надо двигаться в этом направлении.

Закономерности изменения основных параметров и характеристик МОП–транзистора при уменьшении всех линейных размеров в k раз, предсказываемые классической теорией масштабирования, иллюстрируются в приведенной ниже таблице 3.1 (столбец 3).

При постоянном поле необходимо пропорционально уменьшать рабочее напряжение (строка 3) и увеличивать уровень легирования подложки (строка 8), чтобы глубина области пространственного заряда (ОПЗ, англ. *Depletion laer*) (строка 9) также пропорционально уменьшалась. Законы масштабирования остальных указанных в таблице параметров определяются их размерностью, которую можно оценить по формулам, приведенным в столбце 2.

Достаточно очевидно, что уменьшение физических размеров в k раз приводит к пропорциональному уменьшению емкости затвора (строка 5) и рабочего тока (строка 7), а также к соответствующему возрастанию быстродействия (строка 4).

Уменьшается также рассеиваемая каждым элементом мощность (строка 10), однако затрачиваемая на единицу площади мощность при классическом масштабировании не должна меняться (строка 11) в результате увеличения плотности упаковки.

Важным параметром, определяющим производительность вычислений, является энергия, затрачиваемая процессором на операцию с одним битом (англ. *power-delay product*) $E_{\text{bit}} = P \cdot \tau$ (строка 12), которая при классическом масштабировании уменьшается в k^3 раз.

Наконец, при пропорциональном уменьшении всех размеров проводников, соединяющих элементы ИС друг с другом (межсоединений), неизбежно возрастают:

1. сопротивление этих проводников (строка 13) и относительные омические потери в них (строка 16), в результате чего все большая часть подводимой энергии тратится на нагрев соединительных проводников;

2. плотность тока в этих проводниках (строка 15), что предъявляет добавочные требования к их электропрочности.

Таблица 3.1. Законы масштабирования МОП-транзисторов.

Параметр	формула	Класс. F=const	Обобщ. $F=V_D/L$
1	2	3	4
1. Физические размеры L_G, W, d_{ox}, d_{pn}, $d_{ОПЗ}$,межсоединения		$1/k$	$1/k$
2. Электрическое поле	$F = \frac{V_D}{L_G}$	1	$V_D \cdot k$
3. Напряжения V_D, V_G, V_T	$V = F \cdot L$	$1/k$	V_D
4. Время пролета	$\tau = \frac{L}{v}$	$1/k$	$1/k$
5. Емкость затвора	$C_{ox} = \varepsilon \frac{W \cdot L_G}{d_{ox}}$	$1/k$	$1/k$
6. Переносимый заряд	$Q = C_{ox} \cdot (V_D - V_T)$	$1/k^2$	V_D/k
7. Ток	$I = \frac{Q}{\tau}$	$1/k$	V_D
8. Концентрация легирующей примеси в подложке	N	k	$V_D k^2$
9. Глубина области пространственного заряда (ОПЗ)	$d_{ОПЗ} = \sqrt{\frac{2 \cdot \varepsilon \cdot V}{e \cdot N}}$	$1/k$	$1/k$
10.Рассеиваемая мощность	$P = V \cdot I = \frac{C \cdot V^2}{\tau}$	$1/k^2$	V_D^2
11. Плотность мощности на единицу площади	$P_{уд} = \frac{P}{W \cdot L_G}$	1	$(V_D k)^2$
12. Энергия, затрачиваемая на операцию с одним битом	$E_{bit} = P \cdot \tau$	$1/k^3$	V_D^2/k
13. Сопротивление межсоединений	$R_{мс} = \rho_{мс} \cdot \frac{L_{мс}}{S_{мс}}$	k	k
14. Омические потери в межсоединениях	$\Delta V = I \cdot R_{мс}$	1	$V_D k$
15. Плотность тока в межсоединениях	$J = \frac{I}{S_{мс}}$	k	$V_D k^2$
16. Относительные потери в межсоединениях	$\frac{\Delta V}{V}$	k	k

В течение нескольких десятилетий при постоянном уменьшении размеров элементов ИС вплоть до топологической нормы 130 нм все законы классического масштабирования, действительно, удавалось выполнять, хотя это и требовало дополнительных усилий по борьбе с возрастающими паразитными утечками.

Основным критерием возможности увеличения плотности упаковки элементов является мощность, рассеиваемая на чипе. При классическом масштабировании плотность выделяемой на единице площади мощности не должна зависеть от размера элемента (строка 11). Однако, этот параметр учитывает только «активную мощность», затрачиваемую на переключение рабочих элементов. Вместе с тем, при уменьшении размеров резко возрастает и «пассивная» мощность, выделяющаяся при закрытом транзисторе из-за паразитных утечек. Это связано с тем, что при пропорциональном уменьшении напряжения питания разность потенциалов на затворе закрытого (*off*) и открытого (*on*) транзистора также уменьшается по абсолютной величине и, как следствие, экспоненциально возрастают утечки в запертом транзисторе.

В результате для того, чтобы ограничить пассивную мощность при переходе к размерам менее 100 нм, уже не удается пропорционально уменьшать рабочее напряжение, которое остается на уровне 1 В (рис. 3.4). В связи с этим было предложено ввести новые принципы масштабирования, согласно которым напряжение питания VD является еще одной независимой переменной, а обобщенное (англ. generalized) электрическое поле уже не остается постоянным, а определяется формулой F=VD□k. Это, как видно из столбца 4 таблицы, приводит к заметным отличиям в законах масштабирования по сравнению с классикой. Одно из наиболее существенных отличий состоит в том, что рабочий ток (строка 7) уже не уменьшается при масштабировании, а определяется напряжением питания VD независимо от геометрических размеров транзистора. В результате возрастает рассеиваемая на чипе мощность (строка 10), а плотность мощности на единицу площади (строка 11) уже не остается постоянной, а возрастает пропорционально (VD□k)2. Кроме того, при уменьшении размеров существенно замедляется снижение энергии E_{bit}, затрачиваемой на 1 бит (строка 12). Возрастают по сравнению с классикой и омические потери в межсоединениях (строка 14), хотя относительные потери (строка15) остаются такими же.

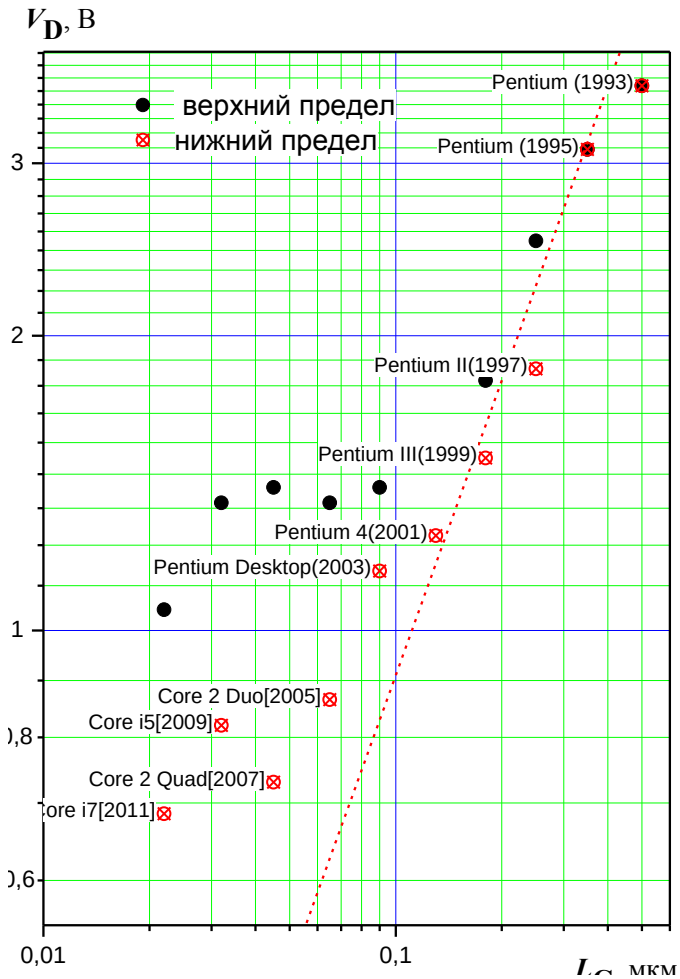

Рис.3.4. Изменение напряжения питания при масштабировании по данным корпорации Intel.

Все это необходимо учитывать при проектировании нового поколения процессоров, тем более, что и быстродействие СБИС, как это видно из рис. 3.1 — тактовая частота не возрастает при уменьшении размеров, — тоже перестает подчиняться законам классического масштабирования по обсуждаемым ниже причинам.

3.2. Фундаментальные ограничения на масштабирование МОП–транзисторов

Очевидно, что экспоненциальный закон уменьшения физических линейных размеров транзисторов (в 1,4 раза каждые 2 года), который успешно выполнялся в течение долгих лет и продолжает пока выполняться в настоящее время, не может продолжаться до бесконечности. Должны быть пределы этому процессу.

За прошедшее время неоднократно высказывались прогнозы об окончании «эры масштабирования». Один из наиболее существенных

был связан с «литографическим порогом», поскольку, согласно общепринятому тогда мнению, из-за дифракционных эффектов невозможно оптическими методами получить пространственное разрешение меньше длины волны излучения (150 нм, т. к. для еще меньших длин волн не существует прозрачных линз). Однако ученые и технологи ценой немалых усилий смогли преодолеть этот барьер.

Другое существенное ограничение связано с тем, что при уменьшении толщины подзатворного окисла SiO_2 до 3 нм и меньше наблюдается катастрофическое возрастание паразитного тока, обусловленного туннелированием электронов из затвора в канал. Однако и эту проблему, как будет показано позже, удалось успешно преодолеть.

Были и другие прогнозы, предсказывавшие, что при уменьшении размеров будет превышено напряжение пробоя кремния, произойдет разрушение межсоединений из-за недопустимо высокой плотности тока, резко возрастет сопротивление канала транзистора и пр. Все эти вопросы к настоящему времени также, в большей или меньшей степени, разрешены, причем работа в данном направлении непрерывно продолжается. А ведущие производители продолжают неуклонно традиции масштабирования размеров активных элементов КМОП.

Тем не менее, фундаментальные ограничения на минимальные размеры МОП–транзисторов все–таки существуют. Первое из них связано с тем, что в любой реализации двоичной логики состояния, соответствующие логической единице, и состояния, определяющие логический ноль, должны быть различимы. Это означает, что для перехода из одного состояния в другое необходимо затратить энергию E_{bit} (энергия для обработки одного бита информации), чтобы преодолеть потенциальный барьер между этими состояниями. В случае классической статистики Больцмана вероятность такого перехода $P_{\text{кл}}$ задается известной формулой $P_{\text{кл}} = \exp\left(-\frac{E_{\text{bit}}}{kT}\right)$.

Состояния будут различимы только в том случае, если эта вероятность меньше (желательно, много меньше), чем 0,5. Отсюда следует очевидное условие различимости состояний:

$E_{\text{bit}} \geq kT \cdot \ln 2$ (0,179 эВ при T = 300 К). 3.1.

Это выражение обычно называется **формулой Шеннона–Неймана–Ландауэра** (*Shannon–Neumann–Landauer*) или просто **SNL.**

Помимо классического перехода электрона над барьером из одного состояния в другое возможно и квантовомеханическое туннелирование между этими барьерами с вероятностью $P_{\text{кв}}$. В этом случае суммарная вероятность обмена электронами между двумя устойчивыми состояниями P_{Σ} определяется законами статистики:

$$P_{\Sigma} = P_{\text{кл}} + P_{\text{кв}} \cdot (1 - P_{\text{кл}}). \qquad 3.2$$

По квазиклассической теории Венцеля–Крамерса–Бриллюэна (ВКБ) — (*Wentzel–Kramer*s–*Brillouin* — **WKB)** — прозрачность $P_{\text{кв}}$ для туннелирования электрона с нулевой начальной энергией сквозь прямоугольный одномерный потенциальный барьер высотой E_{bit} (если его ширина L много меньше длины волны электрона) дается формулой

$$P_{кв} = \exp\left(-4\pi \cdot \frac{\sqrt{2m}}{h} \cdot \int_0^L \sqrt{E_{\text{bit}}}\, dx \right) = \exp\left(-4\pi \cdot \frac{\sqrt{2mE_{\text{bit}}}}{h} \cdot L \right). \qquad 3.3$$

Здесь h — постоянная Планка, а m — эффективная масса электрона.

В первом приближении, положив P_{Σ} равным 0,5, получим условие разрешимости состояний

$$E_{\text{bit}} \geq kT \cdot \ln 2 + \frac{0{,}0045}{L^2}, \qquad 3.4$$

в котором энергия дается в эВ, а длина в нм.

Из формулы 3.4 видно, что при T = 300 К квантовомеханические эффекты могут дать заметный вклад лишь в том случае, когда ширина барьера не больше двух нанометров. Существенно, что второе слагаемое в этой формуле не зависит от температуры, поэтому при достижении таких малых размеров охлаждение не в состоянии понизить рассматриваемый предел.

Более того, поскольку на охлаждение ИС необходимо затратить дополнительную энергию, то суммарные энергетические затраты на обработку одного бита информации (по теореме Карно) не только не уменьшаются, а, напротив, возрастают при понижении температуры:

$$E_{\text{bit}}^{\text{total}} = E_{\text{bit}} + \frac{T_0 - T}{T} E_{\text{bit}} = E_{\text{bit}} \frac{T_0}{T}, \qquad 3.5$$

где T — температура ИС, а T_0 — температура окружающей среды.

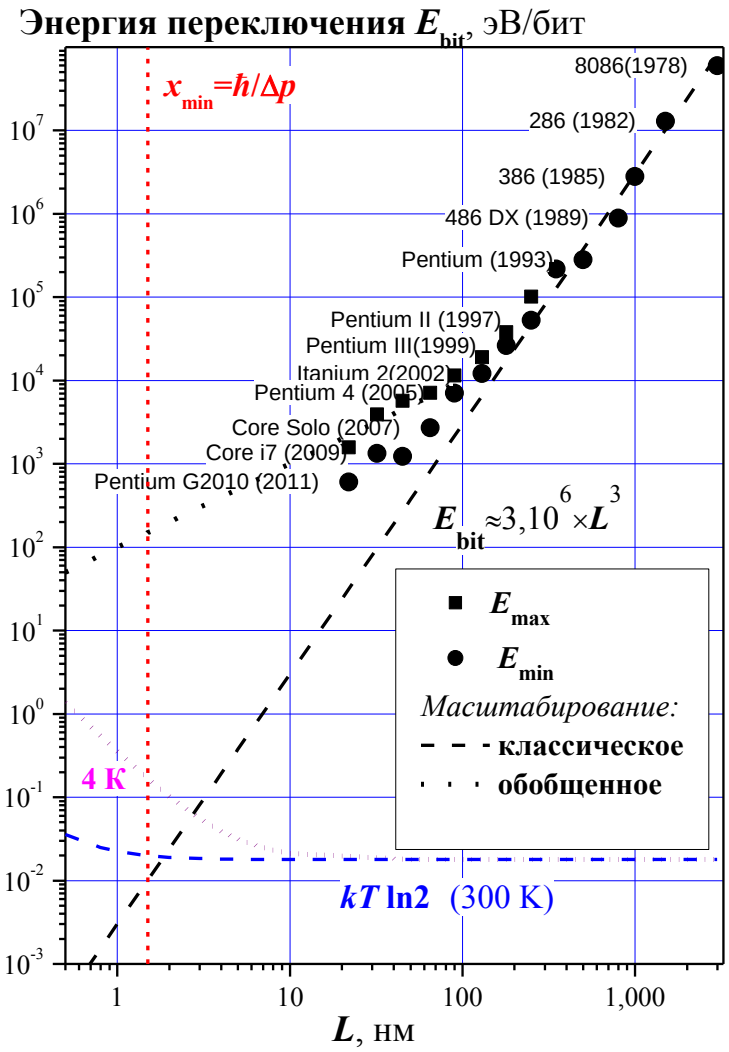

Рис.3.5. Пояснения к фундаментальным ограничениям на масштабирование

За последние тридцать лет при переходе ко все новым проектным нормам энергия E_{bit} стала на пять с лишним порядков меньше, как это иллюстрируется на приведенном рисунке (рис. 3.5).

Точками здесь обозначены расчетные значения E_{bit} для различных топологических норм корпорации Intel с учетом допустимого интервала напряжений питания.

На рисунке также показаны тенденции, предсказанные классической и обобщенной теориями масштабирования. Видно, что при топологической норме ниже 120 нм классическая теория уже не выполняется и энергия E_{bit} спадает при уменьшении размеров по закону, близкому к обобщенной теории, т. е значительно медленнее, чем до этого.

Рассмотренное ограничение на возможные пределы масштабирования, однако, не является главным, так как существуют и другие факторы. В частности, минимально возможные линейные размеры канала транзистора ($x_{\min}$) и максимальная рабочая частота ($f_{\max} = 1/t_{\min}$) ограничиваются известными соотношениями неопределенностей Гейзенберга. В соответствии с ними можно написать:

$$x_{\min} = \frac{h}{2\pi \cdot \Delta p} = \frac{h}{2\pi \cdot \sqrt{2m \cdot E_{\text{bit}}}} = \frac{h}{2\pi \cdot \sqrt{2m \cdot kT \ln 2}} \approx 1{,}5 \text{ нм при } 300 \text{ K} \ . \qquad 3.6$$

$$t_{\min} = \frac{h}{2\pi \cdot \Delta E} = \frac{h}{2\pi \cdot kT \ln 2} \approx \ 4 \cdot 10^{-14} \text{ с при } 300 \text{ K} \ , \qquad 3.7$$

Здесь, как и везде в последующем, для количественных оценок используется значение энергии переключения одного бита E_{bit} при комнатной температуре, предсказываемое классической теорией масштабирования (формула 3.1). Из формулы 3.6 видно, что полученное значение $x_{\min}$ в ближайшее время вряд ли будет ограничивающим фактором, т. к. на порядок меньше современных топологических норм Точно также и предельно возможная частота, определяемая выражением 3.7 — 25 ТГц, — намного больше достигнутой сейчас и пока тоже не является существенным ограничением для технологов.

С другой стороны, еще в 2004 году Патрик Гелсингер (ныне занимающий пост генерального менеджера группы Intel по цифровым корпоративным технологиям), заявил о том, что если продолжать использовать современные методы разработки процессоров, то к 2010 году они будут вырабатывать больше тепла на квадратный миллиметр, чем ядерный реактор. Именно поэтому снижение тепловыделения современных процессоров наравне с повышением их производительности выходит сегодня на первый план.

Действительно, если создать ИС, у которой одновременно будет и самая большая плотность упаковки (определяемая $x_{\min}$ из 3.6) и максимально возможная частота ($t_{\min}$ из 3.7), и оценить, какая при таких условиях должна выделяться мощность P на единицу площади, то получается следующий результат:

$$P = \frac{n_{\max} E_{\text{bit}}}{t_{\min}} = \frac{kT \ln 2}{x_{\min}^2 t_{\min}} \approx \ 3{,}2 \cdot 10^6 \ \frac{\text{Ватт}}{\text{см}^2} \text{ при } 300 \text{ K} \ . \qquad 3.8$$

Для сравнения — у ламп накаливания, которые сейчас активно изымаются из обращения, эта мощность составляет 100 Ватт/см^2, а у

источника жизни на Земле солнца, — «всего» 6000 Ватт/см^2, т. е почти на 3 порядка меньше, чем в выражении 3.8.

Таким образом, уже на современном этапе развития технологии очевидно, что не получится одновременно увеличивать и плотность упаковки элементов, и рабочую частоту — в каждом конкретном случае придется выбирать что – нибудь одно, в зависимости от того, для чего данная ИС предназначена. Все определяется тем, какую максимальную мощность $P_{\max}$ способна отвести от кристалла используемая технология упаковки микросхем, поскольку реально выделяющаяся при работе мощность P должна быть меньше $P_{\max}$:

$$P = k_{\text{eff}} \frac{n_{\max} E_{\text{bit}}}{t_{\min}} \le P_{\max} = 750 \ \frac{\text{Ватт}}{\text{см}^2} \text{ (в перспективе больше)} \ . \quad 3.9$$

Здесь коэффициент $k_{\text{eff}} < 1$ учитывает, что на каждом такте переключается лишь часть транзисторов, и равен отношению их числа к общему количеству активных элементов. Обычное значение этого параметра порядка 0,1–0,01 для CPU и в десять раз меньше для ячеек памяти, но и это ситуацию кардинально не меняет. Кстати, именно по причине обсуждаемых ограничений все последнее время при неуклонном соблюдении закона Мура по периодическому переходу на очередные топологические нормы рабочая частота не увеличивается согласно закону Деннарда, а остается на уровне 2–3 ГГц (рис. 3.1). Увеличению производительности вычислений препятствуют и задержки при обращении к памяти, т. к. время доступа к ней пока не соответствует возрастающим тактовым частотам, а число необходимых для одного обращения тактов резко возрастает при увеличении числа транзисторов в ИС. Так, в начале 90-х годов прошлого века, когда процессоры содержали около миллиона транзисторов (486-е и первые Pentium), для этого требовалось 6–8 тактов. У современных процессоров с миллиардом транзисторов таких тактов требуется уже 224 и более.

В поисках путей дальнейшего повышения производительности микропроцессоров была разработана концепция *многопоточности* — одновременной параллельной обработки нескольких потоков команд. В частности, специалисты Intel создали технологию сверхпоточной обработки данных (HTT — *Hyper-Threading Technology*), которая позволяет одноядерному процессору выполнять параллельно до четырех

программных потоков одновременно. Таким образом, одно физическое ядро центрального процессора (CPU — *central processing unit*) разделено на несколько логических, каждое из которых операционная система определяет как самостоятельный компьютер. Это значительно повышает эффективность выполнения ресурсоемких приложений (например, связанных с аудио– и видеоредактированием, 3D–моделированием, с проведением интенсивных вычислений).

Логическим развитием этой идеи стало создание процессоров с несколькими физическими ядрами на одном кристалле, в которых все ядра работают параллельно и при меньшей тактовой частоте обеспечивают большую производительность, так как параллельно одновременно выполняются несколько независимых потоков инструкций. В принципе, подобный результат можно было бы получить, разместив на одной материнской плате несколько параллельно работающих процессоров. Этот подход, однако, не получил широкого распространения ввиду высокой стоимости как каждого из процессоров, так и самой материнской платы. К тому же в многопроцессорных системах связь между чипами и обращение к памяти осуществляется через сравнительно медленные порты ввода–вывода (англ. *Input–Output* — I/O), что заметно замедляет работу.

Поэтому наиболее эффективным направлением дальнейшего повышения производительности компьютеров большинство производителей считает создание на одном кристалле мультиядерных (англ. *multi–core*) процессоров, содержащих 2, 4, 8, 12, 16 или более независимых параллельно работающих ядер, изготовленных на одном кристалле. Обмен данными между ядрами осуществляется внутри самого чипа без использования медленных I/O портов.

Возможна различная архитектура подобных систем. Например, Intel и AMD — законодатели мод в этом направлении — избрали гомогенный вариант, в котором все ядра процессора одинаковы и могут выполнять одни и те же задачи. Одно из этих ядер выполняет функции центрального процессора (CPU), а остальные тоже являются необходимыми, но их конкретные задачи определяются алгоритмом решаемой проблемы и задаются введенной программой управления.

Напротив, альянс IBM, Sony и Toshiba придерживается гетерогенной архитектуры, в которой ядра процессора выполняют различные задачи. В выпущенном этими фирмами процессоре Cell из девяти ядер одно является процессором общего назначения PowerPC, а восемь остальных — специализированными процессорами, оптимизированными для векторных операций, которые используются в игровой приставке Sony PlayStation 3.

В любом случае каждое ядро процессора управляется независимо и снабжено собственной кэш–памятью, способствующей существенному ускорению его работы.

Первым процессором, предназначенным для массового использования, а не для встроенных систем, стал POWER 4 с двумя ядрами Power PC на одном кристалле, выпущенный компанией IBM в 2001 году. В апреле 2005 года AMD выпустила двухъядерный процессор Opteron, предназначенный для серверов, а через месяц был предъявлен процессор Pentium D фирмы Intel, ставший первым 2-х ядерным процессором, предназначенным для персональных компьютеров. В марте 2010 года появились первые 12-ядерные серверные процессоры Opteron 6100 компании AMD, а через полтора года эта же компания выпустила первые 16-ядерные серийные серверные процессоры Opteron 6200 (кодовое название *Interlagos*).

На сегодня многими производителями процессоров, в частности Intel, AMD, IBM, ARM, дальнейшее увеличение числа ядер в чипе признано как одно из самых приоритетных направлений повышения его производительности.

Более того, корпорация Intel еще летом 2008 года сообщила, что рассматривает возможность интеграции в один процессор нескольких десятков и даже тысяч вычислительных ядер. Такие чипы называют уже не мультиядерными, а многоядерными (*many–core*). Они должны работать с тем же набором инструкций, но содержать в своей структуре мощное центральное ядро и множество вспомогательных ядер, чтобы

Кэш (англ. *cache*, от фр. *cacher* — «прятать») — промежуточный буфер с быстрым доступом, содержащий информацию, которая может быть запрошена с наибольшей вероятностью. Доступ к данным в кэше осуществляется быстрее, чем выборка исходных данных из более медленной памяти или удаленного источника, однако её объем существенно ограничен по сравнению с хранилищем исходных данных.

более эффективно обрабатывать сложные мультимедийные приложения в многопоточном режиме. Кроме ядер «общего назначения», процессоры Intel будут обладать также специализированными ядрами для выполнения различных классов задач — таких, как графика, алгоритмы распознавания речи, обработка коммуникационных протоколов. В процессе работы многоядерные кристаллы Intel должны активизировать только те ядра, которые необходимы для выполнения текущей задачи, тогда как остальные ядра будут отключены. Уже сейчас в техпроцессе используются так называемые «спящие транзисторы», которые отключают питание отдельных блоков памяти при их бездействии, что позволяет в 3 раза уменьшить энергопотребление. Это позволит кристаллу потреблять ровно столько электроэнергии, сколько нужно в данный момент времени, и облегчит условия теплоотвода.

В случае внедрения такой технологии неизбежно придется обновить и операционные системы, и весь банк приложений для того, чтобы оптимизировать их при применении в многоядерной среде и получить максимально возможный выигрыш в производительности и энергопотреблении. Известно, например, что уже сейчас при работе старых однопоточных приложений в мультиядерной среде скорость повышается не более чем на 5%, хотя ожидаемое ее приращение должно было быть намного больше.

3.3. Проблемы, связанные со свойствами материалов и структурой приборов

Традиционная структура МОП–транзисторов, показанная на рис. 3.1, обеспечила уменьшение их линейных размеров от 10 мкм в 70-е годы прошлого века до субмикронного диапазона без внесения каких-либо структурных изменений и в полном соответствии с законами Мура и Деннарда. Это продолжалось в течение многих лет вплоть до достижения проектной нормы 180 нм. Дальнейшее продвижение по этому пути, однако, столкнулось с рядом проблем, для разрешения которых разработчиком пришлось все–таки внести определенные изменения в конструкцию прибора. Рассмотрим причины этого подробнее.

При переходе к очередным проектным нормам уменьшение длины канала транзисторов происходит быстрее, чем снижение рабочего напряжения на стоке (рис. 3.4), поэтому напряженность электрического

поля в канале вблизи стока увеличивается. Если при движении в этом поле энергия, которую электроны набирают на длине свободного пробега, больше отдаваемой в столкновениях, то средняя энергия носителей («электронная температура») возрастает. В сильных электрических полях часть таких горячих электронов может приобрести энергию, достаточную для инжекции в подзатворный окисел через потенциальный барьер (~3,5 эВ). На границе раздела окисел–кремний в процессе изготовления прибора (в основном, во время отжига после металлизации) возникают нейтральные дефекты типа $\equiv$Si-H, в которых одна из четырех связей Si с окружающими атомами замещается атомарным водородом. Горячие электроны разрывают эту слабую связь (~0,3 эВ) и в результате образуются ловушки акцепторного типа с одной незаполненной связью, на которых скапливается отрицательный заряд. Как следствие, происходит неконтролируемое изменение порогового напряжения и возрастание подпорогового тока в запертом транзисторе, а также уменьшение со временем тока стока открытого транзистора из-за того, что заряд на ловушках частично компенсирует создаваемый затвором положительный заряд в подложке. Все это приводит к долговременной нестабильности характеристик и снижению надежности ИС.

Еще один нежелательный эффект в сильных продольных полях связан с генерацией электронно–дырочных пар в *p*–*n*–переходе стока в результате лавинной ударной ионизации. Образовавшиеся дырки, в основном, уходят в подложку и увеличивают ее ток, а электроны пополняют число горячих носителей. Меньшая часть таких дырок, однако, рекомбинирует у истока с инжектированными электронами. В результате вблизи истока возникает дополнительный положительный заряд, снижающий потенциальный барьер истокового *p*–*n*–перехода, который начинает работать как эмиттер биполярного транзистора. Это приводит к дополнительной паразитной утечке между истоком и стоком.

Для борьбы с перечисленными эффектами следует уменьшать напряженность продольного электрического поля в транзисторе, т. е. расширять области пространственного заряда *p*–*n*–переходов. С этой целью в модифицированной конструкции МОП–транзистора области

истока и стока были расширены путем создания менее легированных участков, которые удлиняют их в сторону канала. (рис. 3.6).

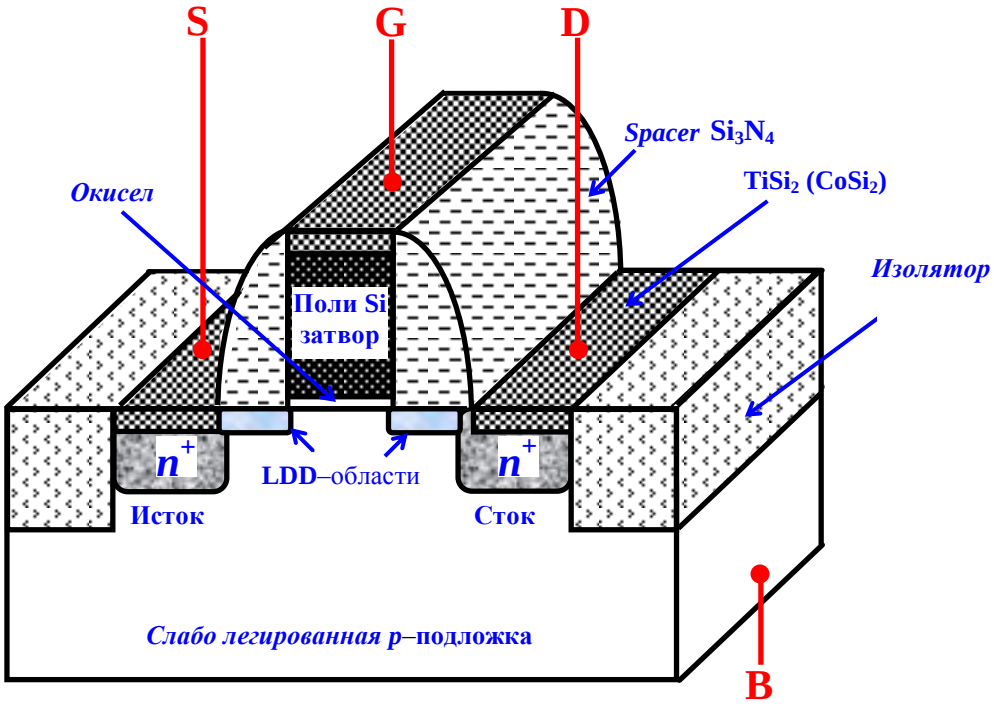

Рис.3.6. Обновленная структура МОП–транзистора

Концентрацию легирующей примеси (P, B) в этих участках и режим её разгонки выбирают таким образом, чтобы получить плавный *p–n–* переход. Обычно концентрация примеси составляет 4–8 10^{18} $см^{-3}$, в то время как в n^+ областях стока и истока она достигает 5–10 10^{19}. Сначала в англоязычной литературе эти области так и назывались — SDE — *Source-Drain-Extension* (расширение стока и истока), однако сейчас во всем мире, включая и Россию, утвердилось название LDD — *Lightly Doped Drain* (слабо легированный сток). Речь идет только о стоке потому, что МОП–транзистор является абсолютно симметричной структурой и не имеет никаких различий между истоком и стоком.

В модифицированную структуру был добавлен еще ряд элементов, отсутствовавших в классической схеме.

1. Начиная с технологических норм 90 нм и менее на характеристики МОП-транзистора начинают влиять соседние элементы топологии ИС, не принадлежащие данному транзистору. Поэтому традиционное использование обратно смещенных *p–n*—переходов уже не обеспечивает требуемой изоляции прибора и приходится вводить специальные изолирующие вставки между соседними приборами.

Первоначально это делалось с помощью техники локального окисления кремния (*LOCOS*), однако вскоре вместо этого стали создавать занимающие меньшую площадь мелкие канавки с окисленными стенками, заполненные поликремнием.

2. При уменьшении размеров потребовалась и более надежная изоляция затвора от стока и истока, которая стала выполняться с помощью специальных разграничителей (спейсеров, англ. *Spacer*) из Si_3N_4, показанных на рис. 3.6.

3. Кроме того при этом возрастают сопротивления всех электродов и контактов к ним, омические потери в них, а также время задержки сигнала из-за увеличения постоянной времени зарядки паразитных емкостей. Для уменьшения роли подобных эффектов между металлом и кремнием стали создавать тонкие промежуточные слои $TiSi_2$ или $CoSi_2$ (рис. 3.6), которые позволили существенно снизить сопротивление контактов ко всем электродам.

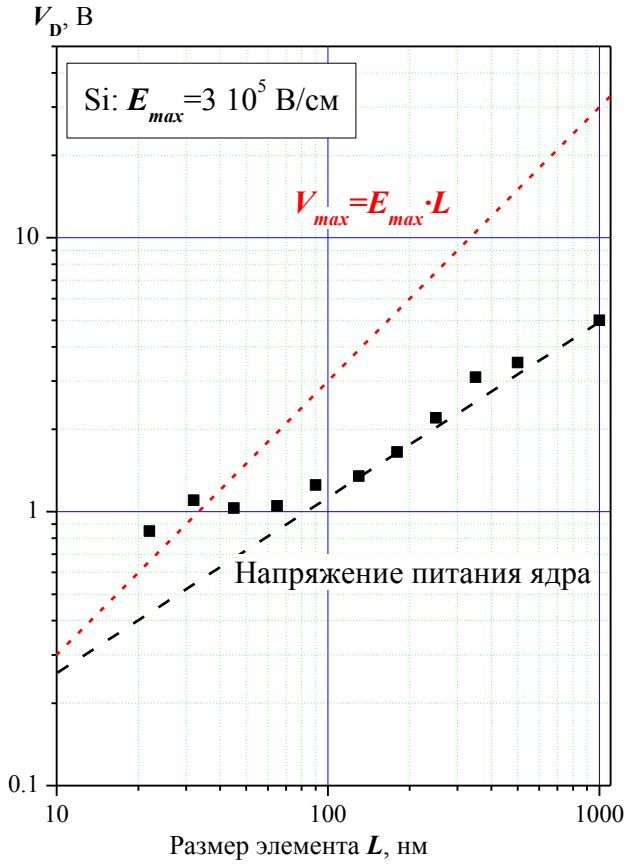

Рис.3.7. Сопоставление среднего напряжения питания и напряжения пробоя Si.

Еще одно ограничение иллюстрируется на рис. 3.7, показывающем, как напряжение питания с введением каждой топологической нормы приближается к максимальному пределу, определяемому электрическим

пробоем в кремнии. В частности, и по этой причине корпорация Intel в 2011 году при введении проектной технологической нормы 22 нм отказалась от показанной на рис. 3.6 конструкции, которая использовалась большинством ведущих производителей в течение десяти с лишним лет, и впервые внедрила принципиально новую трехмерную (3D) структуру МОП–транзистора. Особенности этой технологии подробнее обсудим позднее.