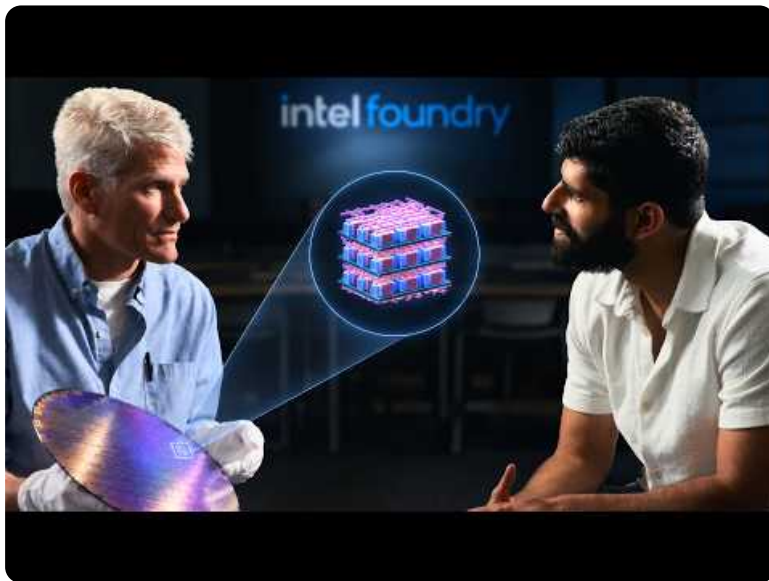


What will computer chips look like in 2045?

Dwarkesh Patel · 2026-10-22 · <https://www.youtube.com/watch?v=frYSOKvNiiE>



В этом выпуске Дваркеш Патель проводит два дня на фабриках (фабах) и в исследовательских лабораториях Intel, чтобы разобраться, как устроено производство самых передовых полупроводников и как физические ограничения железа определяют будущее ИИ. Собеседниками выступают ключевые технические руководители Intel: Крис От (Chris Auth, глава разработки техпроцессов), Кевин Фишер (Kevin Fischer, руководитель интеграции и отладки узлов 18A/18A-P), Джек Кавальерос (Jack Kavalieros, глава исследовательской лаборатории компонентов), Тайлер Осборн (Tyler Osborn, директор по передовой упаковке чипов) и Пушкар Ранаде (Pushkar Ranade, технический директор Intel).

Главная тема выпуска — фундаментальное столкновение экспоненциального спроса со стороны ИИ (который уже превышает максимальные производственные мощности на 40–50%) с жёсткими физическими, технологическими и цепочечными ограничениями кремниевой индустрии. Выпуск охватывает весь стек проблем: от атомной сборки транзисторов с окружающим затвором (Gate-All-Around), литографии High-NA EUV, рутениевых проводников и гибридного бондинга до суровой экономики фабов, монополии поставщиков оборудования и того, почему мир физически не сможет продолжать утраивать ИИ-вычисления каждый год до конца десятилетия.

Разработка техпроцесса 14A и уроки провала 10 нанометров

Масштаб и тонкая настройка современного техпроцесса 01:03: Разработка техпроцесса Intel 14A ведётся уже более трёх лет и сейчас находится на стадии

финализации набора оборудования и тонкой настройки каждого шага. Современный производственный цикл на фабрике включает около 2000 отдельных операций. Каждая из них должна быть откалибрована так, чтобы обеспечивать абсолютную повторяемость при обработке тысяч кремниевых пластин (wafers) в неделю в режиме массового производства (High-Volume Manufacturing, HVM).

Механика поиска дефектов и повышения выхода годных чипов (Yield)

02:07: Ежемесячный рост процента годных чипов (yield) строится на двух уровнях контроля: встроенной метрологии (inline metrology) прямо по ходу конвейера и финальном электрическом тестировании (end-of-line electrical test / sort). Когда электрический тест выявляет сбой, дефектный участок физически разрезают и изучают под просвечивающим электронным микроскопом (TEM) для получения поперечного сечения (cross-section). По срезу определяют, между какими именно слоями возник дефект, вычисляют виновный участок технологического маршрута из 2000 шагов и настраивают там встроенную метрологию для отлова и устранения первопричины.

Катастрофа 10 нм и опасность избыточных инноваций **02:40:** Главный урок задержки 10-нанометрового техпроцесса Intel (планировался на 2016 год, запущен в 2018-м, продукты задержались почти на четыре года, что обрушило акции Intel на 15% и подняло TSMC на 13%) — нельзя закладывать слишком много инноваций в один технологический узел (node). На этапе 10 нм Intel пыталась провести «гипермасштабирование» (hyperscaling) быстрее исторических темпов, причём без новых литографических машин, так как EUV-оборудование было признано ещё сырым. В результате вся нагрузка легла на усложнение самого процесса и инженеров; технология в итоге получилась отличной по стоимости и производительности, но критически опоздала к графику клиентов.

Фактчек: Задержка Intel 10 нм вызвала падение акций Intel на 15% и рост TSMC на 13% — □: Смешаны два события: ограниченные поставки 10-нм Cannon Lake начались в 2018 году, массовые Ice Lake появились в 2019-м, а указанные движения акций относятся к июлю 2020 года и задержке **7-нм** техпроцесса. Claude и GPT согласны с этой поправкой, хотя Claude оценивает ошибку мягче.

Запоздалый переход на EUV и система «фазовых ворот» (Phase Gate)

04:56: Отказ от EUV (экстремальной ультрафиолетовой литографии с длиной волны 13,5 нм, что в ~14 раз короче предыдущих лазеров) на 10 нм стоил Intel лидерства: TSMC внедрила EUV в массовое производство в 2019 году и масштабировала на 5 нм в 2020-м, тогда как первый EUV-процесс Intel дошёл до продуктов только в 2023 году, дав TSMC четырёхлетнюю фору опыта. Чтобы исключить повторение таких срывов, Intel внедрила жёсткую систему метрик зрелости (phase gate): для перехода технологии из стадии поиска (pathfinding) в разработку она обязана пройти строгие пороги по выходу годных, дефектности

и производительности. Если технология (например, 2D-материалы или CFET) не укладывается в метрики к дедлайну, её безжалостно переносят на следующий техпроцесс ради сохранения предсказуемости расписания.

Эволюция транзисторов: от планарных структур к Gate-All-Around

Физика транзистора и проблема утечки тока при уменьшении размеров

06:35: Транзистор работает как переключатель между истоком (source) и стоком (drain), разделёнными кремнием, поверх которого через тонкий слой изолятора лежит затвор (gate). Подача положительного напряжения на затвор притягивает отрицательно заряженные электроны к изолятору, формируя под ним проводящий канал (channel) — переключатель замыкается. По мере миниатюризации исток и сток сблизались настолько, что ток начал пробивать канал даже без напряжения на затворе, вызвав неконтролируемые утечки и потребовав усиления контроля затвора над каналом через переход из 2D в 3D.

От FinFET к Gate-All-Around (GAA) и сложность сборки нанолент **07:41:** В классическом планарном транзисторе (с 1960 года) затвор касался канала только сверху; в FinFET канал превратили в вертикальный «плавник» (fin), что позволило затвору обхватить его с трёх сторон. Следующий шаг — Gate-All-Around (окружающий затвор): канал нарезается на тонкие горизонтальные кремниевые наноленты (ribbons), которые затвор оборачивает со всех четырёх сторон. Чтобы изготовить такие «парящие» ленты, их сначала чередуют со слоями временного жертвенного материала-заполнителя, так как реальные металлы затвора не выдержат последующего отжига пластины при температуре около 1000 °C. После высокотемпературных этапов заполнитель вытравливают, оставляя ленты висеть в воздухе, покрывают их со всех сторон специальным изолятором с высокой диэлектрической проницаемостью (high-k material для усиления притяжения электронов) и заливают металлом затвора.

Почему путь от идеи в лаборатории до конвейера занимает 30 лет **09:20:** И FinFET (продемонстрирован в лаборатории в 1989 году, в серии у Intel с 2011-го), и Gate-All-Around (показан в 1988-м, в серии у Samsung с 2022-го) были изобретены почти одновременно, но внедрялись с разницей в десятилетие по принципу «не решай сложную проблему, пока работают более простые способы». Прежде чем браться за самый сложный GAA, индустрия последовательно выжимала производительность из напряжённого кремния (strained silicon), металлических затворов с high-k диэлектриками (high-K metal gate) и FinFET. Переход к GAA потребовал одновременного изобретения новых химических газов для осаждения, новых нагревательных столиков (heating chucks) и инструментов от внешних поставщиков в долгом итеративном цикле проб и ошибок.

Фактчек: Между ранними демонстрациями FinFET/GAA и серийным внедрением прошли десятилетия — □: Ранние предшественники этих архитектур — DELTA-

транзистор 1989 года и Surrounding Gate Transistor 1988 года. Серийное внедрение Intel Tri-gate в 2011-м и Samsung GAA в 2022-м подтверждает длительный путь от лабораторной идеи до производства, но не устанавливает универсальный срок в 30 лет.

Конец масштабирования Деннарда и десятикратный взрыв числа

материалов 11:35: В 1980–1990-х годах R&D сводился к масштабированию Деннарда (Dennard scaling): достаточно было уменьшать размеры элементов за счёт оптики литографов, а сам технологический маршрут оставался неизменным («главное — ничего не трогать»). В 2000-х этот механизм сломался, открыв «золотой век инноваций», где для каждого нового узла приходится буквально «крушить процесс кувалдой» и менять архитектуру. Количество используемых в производстве химических элементов и материалов выросло в 10 раз по сравнению с 1990-ми, а технологии вроде атомно-слоевого осаждения (Atomic Layer Deposition, ALD) перевели сборку структур на уровень поатомного контроля.

Литография High-NA EUV, подача питания с обратной стороны и переход к Foundry-модели

Экономика и физика машин High-NA EUV 13:11: Для техпроцесса 14A Intel внедряет новейшие литографы ASML High-NA EUV, где числовая апертура (Numerical Aperture, NA — мера ширины конуса света, который зеркала фокусируют на пластине) увеличена со стандартных 0,33 до 0,55. Это позволяет печатать элементы примерно в 1,7 раза меньше за один проход экспонирования. Когда элементы мельче разрешения лазера, рисунок приходится разбивать на 2, 3 или 4 прохода (мультипаттернинг), а каждый лишний проход на машине стоимостью ~200 млн тратит драгоценное время и повышает риск дефектов; High-NA стоит вдвое дороже (~400 млн), но окупается за счёт сокращения числа проходов.

Фактчек: High-NA увеличивает апертуру с 0,33 до 0,55 и позволяет печатать элементы примерно в 1,7 раза мельче — □: Соотношение апертур действительно составляет около 1,67; цены порядка 200 млн за Low-NA и 400 млн за High-NA соответствуют приведённым в ревью оценкам. Однако окупаемость за счёт сокращения мультипаттернинга зависит от конкретного процесса, а в T33 те же \$400 млн уже названы стоимостью комплекса со смежным оборудованием — границы оценки непоследовательны.

Почему High-NA выгодна Intel уже сейчас, а TSMC — нет: роль Backside

Power 14:16: Разное отношение Intel и TSMC к экономической целесообразности High-NA объясняется архитектурным различием в разводке питания. На чипе есть два типа проводов: толстые линии питания (power lines) и тонкие сигнальные линии (signal lines), требующие зазоров для защиты от помех. В прошлом их приходилось втискивать вместе поверх транзисторов, но Intel

перенесла линии питания под транзисторы — на обратную сторону пластины (backside power). Это освободило место сверху и позволило расслабить шаг (pitch) сигнальных линий до уровня, который High-NA печатает за один проход; у TSMC нет подачи питания с обратной стороны, поэтому их шаг линий сверху слишком плотный даже для High-NA, им всё равно приходится делать двойной проход (double patterning) на обычном EUV, из-за чего переплата за High-NA не даёт им выигрыша.

Фактчек: У TSMC нет backside power, поэтому High-NA ей невыгодна — ▲:

Отсутствие обратного питания относится к N2, но не ко всей дорожной карте TSMC: для A16 предусмотрена Super Power Rail, с заявленным серийным выпуском в 2026 году. Экономика High-NA зависит также от проектных правил, конкретного слоя, производительности сканера и стоимости альтернативного паттернинга; Gemini и GPT отвергают однофакторное объяснение, тогда как Claude допускает его текущую применимость.

Болезненная трансформация Intel из IDM в контрактного производителя (Foundry) и проблема PDK 16:26:

Десятилетиями Intel работала как IDM (Integrated Device Manufacturer), производя чипы только для себя: внутренние команды обменивались сырыми данными о транзисторах напрямую без формализованных пакетов проектирования (Process Design Kit, PDK — свод правил и меню доступных элементов техпроцесса для внешних дизайнеров). Из-за отсутствия культуры создания внешних PDK на узле 18A возникли публичные проблемы и задержки при работе со сторонними клиентами. Собрав обратную связь, Intel внедрила строгие стадии зрелости (PDK 0.5, 0.9, 1.0) и заложила высококачественный внешний PDK в узел 14A с самого старта (при этом использование High-NA на 14A оставлено как выбираемая опция внутри PDK под конкретного заказчика).

[18:00-19:15] Рекламная интеграция сервиса корпоративной авторизации WorkOS.

##Взгляд в 2045 год: 3D-стек, CFET и тепловой предел

Альтернативные ветки технологического дерева и будущее 3D-интеграции

19:15: Структура транзистора Gate-All-Around была физически неизбежной вершиной эволюции, которую предвидели ещё в 1990-х, тогда как полное спиливание кремниевой подложки ради прокладки питания снизу (backside power) казалось тогда фантастикой. Сочетание GAA и питания снизу переводит индустрию в новый режим: транзистор теперь буквально «парит» внутри слоёв межсоединений (interconnect). Это открывает путь к монолитному 3D-стекированию к 2045 году — чередованию слоёв транзисторов и проводов прямо при выращивании одного кристалла (подобно тому, как уже строится 3D NAND и куда движется DRAM), хотя горизонт чёткого инженерного планирования в индустрии всегда ограничен лишь четырьмя годами.

Транзисторы CFET и фундаментальный барьер отвода тепла 21:41:

Практическим шагом к 3D станет архитектура CFET (Complementary Field-Effect Transistor), где транзисторы n-типа и p-типа (работающие в паре почти в каждой схеме с противоположными зарядами) строятся не рядом друг с другом, а друг над другом, что в теории удваивает плотность на ту же площадь. Однако логические чипы потребляют гораздо больше энергии, чем память, поэтому их нельзя складывать в многослойные стопки так же легко, как HBM или 3D NAND. Вычисления всегда ограничены тремя факторами: частотой переключения (frequency), энергоэффективностью переключения и отводом выделяемого тепла (thermals); именно перегрев запертых внутри 3D-структуры слоёв становится главным физическим лимитом, требующим создания новых инструментов системного теплового моделирования.

Культура принятия решений на фабрике: совещания Hot Topics и система HotBox

Принятие риска в условиях неполных данных на совещаниях Hot Topics

23:54: На ежедневных кризисных совещаниях Hot Topics по доводке узлов 18A и 18A-P руководство в жёстком режиме распределяет ограниченные ресурсы инженеров и экспериментальных пластин между тысячами проблем. Рядовые инженеры часто не чувствуют финансовых ставок для компании и останавливаются, получив отказ на запрос ресурса, либо пытаются собрать идеальную статистику. Кевин Фишер подчёркивает главный закон индустрии: полного набора данных не будет никогда, и если ждать полной уверенности, время уже упущено — поэтому задача руководителя сводится к ежедневному балансированию риска и принудительному ускорению экспериментов.

Борьба с эпитаксиальными наростами (Epi Nodules) 25:50: Пример реальной проблемы на 18A — образование аномальных наростов (epi nodules) в процессе эпитаксии (выращивания кристаллических структур истока и стока по бокам затвора путём пропускания газов над нагретой пластиной, где кристаллическая решётка подложки служит матрицей для поатомной укладки). Иногда кристалл начинает расти аномально быстро, как «опухоль», замыкаясь на соседний элемент и выводя транзистор из строя. Поскольку заранее неизвестно, какое решение сработает быстрее, команды запускают параллельную разработку сразу по всем направлениям: одни создают специальную химическую очистку для удаления уже выросших наростов, а другие ищут и устраняют фундаментальный источник их появления.

[26:35-27:41] Рекламная интеграция трейдинговой компании Jane Street и их инструмента отладки magic-trace на базе Intel Processor Trace.

Приоритетные лоты HotBox и компромисс скорости против КПД фабрики

27:41: Чтобы критический эксперимент не ждал стандартного цикла обработки

пластины (который длится месяц и более), на фабрике действует многоуровневая система ускорения лотов — HotBox. Высший приоритет, HotBox 1, заставляет партию кремния «нестись по фабрике», получая внеочередной доступ к каждому станку. Производственники ненавидят использовать HotBox 1, так как это работает подобно выделению пустой полосы на шоссе под одну VIP-машину: станки простаивают в ожидании приоритетного лота, что снижает общую эффективность (утилизацию) всей фабрики ради мгновенного получения экспериментальных данных.

Внутри R&D-лаборатории: рутениевые соединения, CFET и отсев идей

Предел медных проводников и переход на рутений на шаге 20 нм 29:30:

Над транзисторами современного чипа нагромождено более дюжины слоёв медных проводов (interconnect). При нанесении меди обязательно требуется подслой — лайнер (liner), работающий как «клей» и барьер для удержания меди. Лайнер обладает высоким сопротивлением, и по мере сужения проводов он занимает всё большую долю сечения, съедая место для самой меди. Intel исследует замену меди на рутений (ruthenium): хотя у рутения собственное удельное сопротивление выше, он поддаётся прямому травлению (etching) и вообще не требует лайнера. Точка пересечения, где рутений без лайнера выигрывает по проводимости у меди с лайнером, находится в районе шага линий (pitch) в 20 нанометров (10 нм металл и 10 нм диэлектрик-изолятор).

Фактчек: Рутений может выигрывать у меди с барьерным слоем при шаге межсоединений около 20 нм — □: На таком масштабе барьер и лайнер занимают значительную долю сечения медного проводника, поэтому более высокое объёмное сопротивление рутения может компенсироваться преимуществами интеграции. Рубеж около 20–25 нм поддержан приведёнными Gemini исследованиями, но это технологически зависимая оценка, а не универсальная граница для любых конструкций.

Урок кобальта на 10 нм и возвращение меди 31:36: Выбор материала определяется не только его физикой в изоляции, но и способностью интегрироваться в экосистему с внешними поставщиками оборудования. На 10-нанометровом узле Intel переходила с меди на кобальт, так как на момент принятия решения кобальт показывал лучшие результаты. Однако позже совместная работа с вендорами позволила радикально утончить лайнер для меди, что вернуло медь на лидирующие позиции в текущих узлах — до тех пор, пока дальнейшее сжатие не вынудит окончательно перейти на рутений.

Сборка CFET в лаборатории: проблема вертикальных траншей и «короткие циклы» (Short-Loop Vehicles) 32:45: Главная сложность вертикального наложения транзисторов в CFET — не само наложение, а вертикальная разводка контактов между верхним и нижним этажами. Это

требует вытравливания чрезвычайно глубоких и узких колодцев с высоким аспектным отношением (aspect ratio — отношение глубины к ширине) и их последующего заполнения металлом без образования воздушных пустот (air gaps). Чтобы не ждать по месяцу прогона полного цикла из 2000 шагов, лаборатория использует «короткие циклы» (short-loop vehicles): сначала тестируются простейшие базовые ячейки (диоды, конденсаторы), затем упрощённый CFET с одной нанолентой NMOS поверх одной ленты PMOS (вместо четырёх на четыре), и лишь после решения проблем с травлением и химией конструкция усложняется до релизной.

Откуда берутся 10 лет задержки между лабораторным образцом и смартфоном 35:52: Уже сегодня в лаборатории Intel с помощью EUV и электронно-лучевой литографии (E-beam) печатают структуры финальных размеров и осаждают двумерные (2D) материалы толщиной в один атомный монослой (дихалькогениды переходных металлов). Однако превращение работающего одноленточного прототипа в коммерческий чип занимает около 10 лет из-за «длинного хвоста» интеграции: нужно нарастить число лент, подобрать совместимые диэлектрики и идеально состыковать более 1000 технологических шагов так, чтобы новая архитектура превзошла старую по всем параметрам одновременно.

Почему 90% новых материалов отбраковываются: адгезия и газовые реакции 38:25: Большинство перспективных идей (например, из пула в 20 кандидатов 2D-материалов) погибают при столкновении с суровой реальностью смежных шагов конвейера. На базовом уровне материал может отлично наноситься и проводить ток, но отсоединиться из-за банального отсутствия адгезии — к нему просто ничего не прилипает на следующем шаге создания контакта. На более сложном уровне материал может не выдерживать последующей обработки: если покрывающий его защитный слой (cap) не абсолютно герметичен, газы на четвёртом или пятом шаге вниз по конвейеру проникают внутрь и необратимо меняют его электрические свойства.

Передовая упаковка (Advanced Packaging) и гибридный бондинг

Проблема «площадь против периметра» и эволюция упаковки до 2.5D 39:45: В современных ИИ-ускорителях (например, Nvidia Blackwell с двумя огромными вычислительными кристаллами и восемью стеками памяти HBM вокруг) узким местом является не скорость вычислений, а скорость пересылки данных между чипами. Если удвоить длину и ширину квадратного кристалла, его вычислительная площадь вырастет в 4 раза, а периметр краёв, через которые идут провода к соседям, — только в 2 раза. Чтобы обойти этот геометрический предел, индустрия перешла от размещения чипов рядом на плате к 2.5D-упаковке — установке нескольких кристаллов на общую кремниевую подложку-посредник (interposer) с плотной внутренней разводкой.

Принцип работы гибридного бондинга (Hybrid Bonding) 41:25: Следующий скачок плотности — 3D-наложение чипов друг на друга. Ранние версии использовали микроскопические шарики припоя (solder bumps), но они слишком громоздкие и создают высокое электрическое сопротивление. Гибридный бондинг полностью убирает припой: медные контактные площадки двух чипов с высочайшей точностью совмещаются лицом к лицу, чипы прижимаются и нагреваются, в результате чего медь расширяется и сплавляется напрямую через зазор, а окружающие её диэлектрические поверхности намертво спекаются друг с другом.

[42:25-43:50] Рекламная интеграция платформы детерминированного тестирования кода Antithesis.

Дилемма гибридного бондинга: Wafer-to-Wafer против Chip-to-Wafer 43:50:

Существует два метода гибридной стыковки: наложение целой пластины на пластину (wafer-to-wafer, доминирует в производстве CMOS-сенсоров камер и тестируется для памяти) и наложение отдельных нарезанных кристаллов на пластину (chip-to-wafer / die-to-wafer). Проблема wafer-to-wafer в том, что кристаллы должны быть идентичны по размеру, а случайное распределение бракованных чипов на двух пластинах приводит к тому, что годный чип сверху часто приклеивается к мёртвому чипу снизу, убивая оба. В методе chip-to-wafer пластину сначала режут, отбирают гарантированно рабочие кристаллы (known good dies) и приваривают их только к проверенным живым позициям на базовой пластине — именно этот подход наиболее разумен для стыковки дорогих логических чипов.

Сжатие шага контактов в 300 раз и стирание границы между чипом и упаковкой 45:30: Пятнадцать лет назад шаг контактов первого уровня упаковки составлял 130–150 микрометров; сегодня в серийной технологии Intel Foveros он равен 25 мкм, в лаборатории — менее 10 мкм, а на конференции IEEE ECTC уже продемонстрирована надёжная работа гибридного бондинга с шагом 0,45 мкм. При этом внутри самого кремниевого кристалла верхние слои межсоединений имеют шаг как раз от 0,5 до 1 мкм. Выход упаковки на уровень $\leq 0,5$ мкм означает, что с соединяемых кристаллов можно будет вообще убрать верхние слои металлизации и напрямую спаивать внутренние слои двух разных чипов со снижением падения напряжения (IR drop) и энергопотребления (в пикоджоулях на бит), что даст огромный выигрыш даже в клиентских ноутбуках за счёт экономии батареи и снижения нагрева.

Глобальный дефицит оборудования, экономика масштаба и гонка с TSMC

Тотальный дефицит оборудования и жёсткая привязка к вендорам (Vendor Lock-in) 48:05: Узким горлышком для масштабирования выпуска чипов

под нужды ИИ является не только ASML, производящая менее 100 EUV-сканеров в год без планов радикального роста до конца десятилетия (которые делят между собой Intel, TSMC и производители памяти вроде SK Hynix). Острый дефицит наблюдается по всему спектру фабричного оборудования. Даже если у нелиитографического станка есть несколько поставщиков на рынке, после фиксации техпроцесса (например, 18A) фабрика оказывается намертво привязана к конкретной модели конкретного вендора и не может заменить её аналогом без переделки технологии; поэтому при переходе с 18A на 14A Intel проектирует процесс так, чтобы переиспользовать около 80% существующего парка машин.

Фактчек: ASML выпускает менее 100 EUV-сканеров в год и не планирует существенно наращивать выпуск до конца десятилетия — Δ : Текущий масштаб выпуска менее 100 машин поддерживается ревью, но отсутствие планов роста — отдельное и спорное утверждение. Gemini приводит сведения о расширении мощностей в 2027–2028 годах, однако ссылок в предоставленных материалах нет: точные проценты и срок превышения 100 машин требуют проверки по первичным сообщениям ASML.

Почему мир не сможет продолжать утраивать ИИ-вычисления каждый

год 50:05: На существующем парке оборудования выжать дополнительный объём пластин можно лишь минимально, так как выход годных (yield) уже высок, а каждый дополнительный 1% даётся с огромным трудом. Даже при бесконечных деньгах на капитальные затраты (CapEx) в мире физически существует достаточно станков, чтобы увеличить производство передовых кремниевых пластин максимум в 2 раза (2x), но никак не в 10 раз (10x). Это означает, что десятилетний тренд ИИ-индустрии по ежегодному утроению (3x в год) вычислительных мощностей неизбежно упрётся в физическую стену, если вся цепочка поставщиков оборудования и их субподрядчиков синхронно некратно нарастит выпуск.

Фактчек: Производство пластин может вырасти максимум вдвое, поэтому ежегодное утроение ИИ-вычислений неизбежно прекратится — Δ : Потолок 2x не подкреплён сроком, расчётом оборудования и независимой оценкой; число пластин нельзя напрямую приравнять к вычислениям, которые растут также благодаря архитектуре, плотности, точности операций и загрузке ускорителей. Gemini принимает этот потолок как основу прогноза, тогда как Claude и GPT считают его недоказанной оценкой и различают рост отдельных обучающих запусков и всего мирового парка.

Перевернутая пирамида ценности и возврат к псевдо-вертикальной

интеграции 51:07: В экономике ИИ возник поразительный мультипликатор: пара миллиардов долларов CapEx на стороне Intel (амортизируемых годами) создаёт логические чипы, способные генерировать около \$100 млрд ежегодной выручки для ИИ-лабораторий на вершине стека. На протяжении 30–40 лет полупроводниковая индустрия переходила от вертикальной интеграции к

горизонтальной специализации, но сегодня гиперскейлеры и ИИ-компании сами проектируют чипы, обладают триллионными капиталами и генерируют спрос, способный целиком поглотить мощность одной или нескольких фабрик. Индустрия возвращается к «виртуальной вертикальной интеграции», где ИИ-гиганты готовы напрямую финансировать постройку фабов ради гарантированных квот кремния.

Жестокая экономика масштаба: почему в мире осталось всего 3

передовых фаундри 53:17: Полупроводниковое производство — чистейший пример экономики масштаба, так как после постройки фабрики около 80% структуры затрат являются фиксированными. В таких условиях каждый дополнительный процент загрузки станка (utilization) или роста выхода годных чипов (например, с 60% до 70%) даёт нелинейный взрывной прирост маржинальности с той же площади. Именно неподъёмность фиксированных костов привела к вымиранию производителей: если 40 лет назад передовых фабрик было около 20, то сегодня в гонке остались только три (TSMC, Samsung, Intel), причём TSMC отгружает в 3 с лишним раза больше передовых пластин, чем Intel.

Мастер-план Intel по сокращению отставания от TSMC: тайминг важнее

пары процентов плотности 55:07: Хотя передовой узел TSMC N3 полностью перегружен заказами на ИИ-ускорители (в основном от Nvidia), а бум ИИ параллельно вызвал гигантский дефицит серверных CPU, догнать TSMC по физическому объёму мощностей можно только постепенно (сейчас годовой CapEx TSMC составляет ~50 млрд против ~20 млрд у Intel). Однако технологически, по словам СТО Intel Пушкara Ранаде, паритет уже почти достигнут: разница в несколько процентов по плотности или частоте между Intel 14A и TSMC N2 не имеет решающего значения, так как техпроцесс не одномерен — он включает целое семейство транзисторов (с низкими утечками, с высоким током), разные типы памяти SRAM (высокоплотную и высокоскоростную), аналоговые и радиочастотные (RF) блоки. Для клиентов гораздо важнее предсказуемость сроков (on time) и высокий начальный yield для быстрого масштабирования, чем выигрыш в 2–3% площади кристалла.

Фактчек: TSMC N3 загружен преимущественно Nvidia, а Intel 14A почти равен TSMC N2 — △: Эти утверждения не подтверждены сопоставимыми данными: Blackwell производится на TSMC 4NP, а среди крупных потребителей N3 есть Apple, поэтому необходима привязка загрузки к конкретной дате. Паритет 14A/N2 нельзя устанавливать без сравнения плотности, мощности, быстродействия и выхода годных; важность сроков поставки сама по себе такого паритета не доказывает.

Продажа бизнеса памяти Intel и инерция цепочек поставок до 2030 года

59:10: Историческая ирония заключается в том, что Intel начинала как производитель памяти DRAM в конце 1960-х, а несколько лет назад продала свой бизнес флеш-памяти NAND — ровно перед тем, как ИИ превратил память в

многоотриллионный дефицитный рынок, где нехватка мощностей сегодня даже острее, чем в логических чипах. Из-за жёстких временных лагов (срок поставки заказанных сегодня станков составляет 18 месяцев, что сдвигает их получение к 2028 году, плюс время на постройку корпуса фабрики) глобальное производство чипов к 2030 году не сможет существенно превысить текущие прогнозы аналитиков.

Фактчек: Intel продала NAND-бизнес перед бумом памяти для ИИ — □: Продажа SK hynix за \$9 млрд была объявлена в 2020 году и завершена в 2025-м. Однако NAND для SSD и DRAM для HBM — разные технологии: сохранение NAND-бизнеса само по себе не обеспечило бы Intel участие в буме HBM.

Фактчек: Память уже стала многоотриллионным рынком, а сроки поставки станков фиксируют потолок производства до 2030 года — □: Как утверждение о текущей годовой выручке «многоотриллионный рынок» не подтверждается: Claude приводит порядок \$150–250 млрд для DRAM и NAND в 2025 году, а GPT также отвергает триллионную оценку. Gemini подтверждает исходную формулировку, но приведённые им показатели отдельной Solidigm не доказывают масштаб всего рынка; срок поставки оборудования в 18 месяцев также недостаточен для вывода о жёстком потолке до 2030 года.

Terafab Илона Маска и радикальное переосмысление фабрик будущего

Партнёрство с Илоном Маском по проекту Terafab 1:01:07: Желание Илона Маска построить собственную гигантскую фабрику чипов (Terafab) продиктовано тем же осознанием грядущего голода на кремний для вычислений и памяти внутри его компаний. Поскольку запуск полупроводникового производства несоизмеримо сложнее автомобильного завода, Intel сотрудничает с Маском, чтобы ускорить развёртывание этих мощностей; при этом на фоне дефицита станков строительство новых фабов всё равно необходимо для расширения чистого физического пространства (clean room capacity) и подведения колоссальных объёмов электричества, воды и расходных материалов.

Фактчек: Intel официально сотрудничает с Маском по Terafab — △: Здесь сохраняется расхождение в подтверждённости: Gemini сообщает об официальном присоединении Intel в апреле 2026 года и приводит параметры проекта, Claude прямо указывает на отсутствие подтверждения. Без первичного объявления в предоставленных материалах нельзя считать установленными ни статус партнёрства, ни бюджет, ни сроки запуска.

Закон удвоения стоимости фабов и абсурдность сканеров за 800 млн

1:02:48: На протяжении последних 30–40 лет стоимость постройки полупроводниковой фабрики удваивается каждые 4–5 лет, а стоимость самого дорогого модуля — литографии — растёт ещё быстрее. Комплекс High-NA EUV со смежным оборудованием уже стоит около 400 млн; линейная экстраполяция

означает, что через четыре года инструмент паттернинга будет стоить \$800 млн, что разрушает экономический смысл производства и требует радикального пересмотра базовых догм.

Фактчек: Удвоение стоимости фабрик означает появление сканера за \$800 млн через четыре года — □: Это сценарная экстраполяция, а не неизбежное следствие исторической тенденции: стоимость фабрики и отдельного сканера не обязана расти одинаково. Повторяющееся удвоение является экспоненциальным, а не линейным ростом; экономическую состоятельность определяет прежде всего стоимость годного кристалла.

Индоктринация индустрии: а нужна ли современному фабу чистая комната? 1:04:10: Из-за того, что на кону стоят десятки миллиардов долларов, а день простоя приносит колоссальные убытки, полупроводниковая отрасль патологически боится риска и десятилетиями копирует одни и те же подходы, превращая их в религиозную догму для новых поколений инженеров (то же самое касается и дизайна CPU, базовые принципы которого застыли 30 лет назад). Пушкар Ранаде приводит мысленный эксперимент: учитывая, насколько герметизированы современные процессы внутри самих станков, действительно ли сегодня нужны гигантские и безумно дорогие чистые комнаты (clean rooms)? Можно было бы провести эксперимент, запустив в исследовательский фаб людей с гамбургерами и сигарами, чтобы проверить реальное влияние на yield, и потенциально сэкономить миллиарды на инфраструктуре — но ни один гигант (Intel, TSMC, Samsung) никогда добровольно не пойдёт на такой риск без подхода «с чистого листа».

Стратегический вывод: необходимость агрессивной скупки оборудования 1:07:36: Поскольку главным ограничителем дохода в ближайшие годы станут не стены фабрик, а квоты на станки от поставщиков верхнего уровня, оптимальная стратегия для Intel — привлекать максимум капитала и заранее выкупать обязательства по поставкам оборудования (подобно тому, как Nvidia законтрактовала память на сотни миллиардов долларов). Признавая, что с учётом сегодняшних знаний инвестировать агрессивнее стоило ещё год назад, руководство Intel намерено наращивать капитальные вложения в закупку машин, чтобы удержать передовое производство внутри США.

ЛИЧНО ДЛЯ ТЕБЯ

Не привязывай бюджет на ИИ к одному прогнозу дефицита: Ограничения оборудования и памяти реальны, но не гарантируют сохранения высоких цен API — эффективность моделей и конкуренция тоже влияют на стоимость. Сравнивай Claude, Codex/ChatGPT и Gemini по цене успешно выполненной задачи, включая время на проверку результата.

Локальное AI-железо выбирай по памяти и своему сценарию: Для генерации с небольшим батчем объём и пропускная способность памяти часто важнее пиковых FLOPS. Перед покупкой проверяй нужную модель, квантование, длину контекста и фактическую скорость, а не только рекламные показатели.

На стройке ограничивай число одновременных новшеств: Для нестандартных инженерных решений заранее назначай срок проверки, критерии готовности и проверенный запасной вариант. Если решение не прошло проверку вовремя, не позволяй ему автоматически задерживать весь дом.

Сокращай цену неопределённости короткими проверками: Перед крупной тратой выделяй самое рискованное предположение и проверяй его отдельно — образцом материала, замером или небольшим испытанием. Решение о продолжении принимай по заранее заданным критериям, а не после бесконечного накопления информации.

КОНСЕНСУС РЕЦЕНЗЕНТОВ

Ревью сходятся в значимости производственных ограничений и длительных циклов внедрения; Claude и GPT также согласны в исправлении истории 10 нм и критике прямого равенства «пластины = ИИ-вычисления». Главные расхождения — степень доверия к потолку роста 2х, оценке рынка памяти и свежим сведениям о Terafab; планы расширения ASML приведены только Gemini и требуют отдельного подтверждения.