

Dylan Patel — The single biggest bottleneck to scaling AI compute

@DwarkeshPatel · 2026-03-13 · https://www.youtube.com/watch?v=mDG_Hx3BSUE



Выпуск подкаста Дваркеша Пателя с Диланом Пателем, CEO аналитической компании SemiAnalysis. Разговор посвящен реальным физическим, экономическим и логистическим ограничениям на пути к масштабированию искусственного интеллекта и достижению AGI.

Главная мысль выпуска: вопреки популярному нарративу, финальным препятствием для масштабирования ИИ станут не нехватка электроэнергии или земли для дата-центров, а физические ограничения цепочек поставок полупроводников. В первую очередь — производство литографических EUV-установок компанией ASML и глобальный дефицит мощностей для производства памяти, который радикально изменит рынок потребительской электроники.

Экономика вычислений и стратегии ИИ-лабораторий

Капитальные затраты Big Tech и ИИ-лабораторий 00:31: Совокупный CapEx «Большой четверки» (Amazon, Meta, Google, Microsoft) в этом году составит 600 млрд, а с учетом всей цепочки поставок цифра доходит до 1 трлн. В США в этом году будет введено 20 ГВт новых мощностей. Недавние раунды финансирования OpenAI (110 млрд) и Anthropic (30 млрд) полностью покрывают их текущие затраты на аренду вычислений, которые составляют \$10-13 млрд за 1 ГВт в год.

Фактчек: CapEx «большой четвёрки» ~600 млрд в 2026, раунды OpenAI 110 млрд и Anthropic 30 млрд — □: Гайдены на 2026 (Amazon ~200, Google ~180, Microsoft ~140+, Meta ~115-135 млрд) дают ~650 млрд; раунды совпадают с публичными цифрами. Gemini сначала отнесла \$600 млрд к 2024 году и назвала

завышением в 3 раза, но для 2026 сама признаёт цифру верной — расхождение только в датировке.

Дефицит мощностей Anthropic 02:51: Anthropic необходимо достичь мощности в 4-5 ГВт к концу года только для того, чтобы поддерживать текущий рост выручки (добавление 6 млрд ARR ежемесячно). Из-за изначальной консервативности в закупках Anthropic упустила возможность заключить долгосрочные контракты по низким ценам, в отличие от OpenAI. Теперь компания вынуждена арендовать мощности у провайдеров второго эшелона, покупать спотовые инстансы с огромной наценкой (до 2,40 в час за H100 при себестоимости \$1,40) или отдавать долю выручки облачным платформам (Bedrock, Vertex, Foundry).

Опровержение быстрой амортизации GPU 10:41: Тезис Майкла Бьюрри о том, что GPU обесцениваются за 3 года из-за выхода новых поколений, неверен. Стоимость чипа определяется не его характеристиками, а ценностью извлекаемых из него токенов. Модель GPT-5.4, работающая на H100, генерирует кратно больше экономической ценности, чем GPT-4 на том же чипе. Из-за этого H100 сегодня стоит дороже, чем три года назад. В случае создания AGI один чип H100 сможет окупать себя за несколько месяцев, заменяя труд белых воротничков.

Фактчек: H100 дороже, чем три года назад; Бьюрри неправ про 3-летнюю амортизацию — Δ : Аренда H100 упала с ~8/ч (2023) до ~2/ч (2025), на дефиците конца 2025 отскочила, но не вернулась к уровню 3-летней давности. Тезис «ценность чипа = ценность токенов» здравый, но он продлевает полезность GPU, а не отменяет экономическое обесценение. Бьюрри спорил не с «3 годами», а с растягиванием амортизации гиперскейлерами до 5-6 лет.

Эффект Алчиана-Аллена в ИИ 19:42: Рост фиксированной стоимости аренды GPU (например, с 2 до 3 в час) сокращает относительную разницу в цене между топовыми и средними моделями. Если генерация токенов Opus стоит дороже Sonnet, то при удорожании базового железа клиентам становится выгоднее переплачивать за самую умную модель. Это позволяет облачным провайдерам и производителям чипов (Nvidia, TSMC, производители памяти) забирать себе всю маржу, фиксируя сверхприбыли через долгосрочные контракты.

Фактчек: Эффект Алчиана-Аллена: рост цены GPU-часа делает топ-модель относительно дешевле — Δ : Расхождение. GPT — $\square$: эффект требует одинаковой абсолютной надбавки к дешёвому и дорогому товару, а рост цены GPU-часа сильнее бьёт по модели, которая жжёт больше GPU-часов, и относительные цены сам по себе не сближает; «поставщики забирают всю маржу» из этого тем более не следует. Gemini и Claude приняли механизм без проверки. Практический вывод («не экономь на модели») может быть верен, но по другой причине — из-за стоимости твоего времени.

Битва за мощности TSMC и ошибка Google

Распределение 3-нм мощностей TSMC 26:15: Исторически TSMC отдавала приоритет производителям CPU (Apple, AMD, Amazon Graviton), считая этот рынок более стабильным. Однако Nvidia забрала львиную долю 3-нм мощностей под архитектуру Rubin, потому что раньше всех осознала масштабы спроса и внесла огромные невозвратные депозиты. Apple постепенно вытесняется на 2-нм техпроцесс, объемы которого также будут сокращаться из-за падения спроса на смартфоны.

Информационная асимметрия между Google и Anthropic 31:19: В третьем квартале прошлого года Google Cloud продала Anthropic доступ к 1 миллиону чипов TPU v7 (Ironwoods). Руководство Google не предвидело взрывного роста ИИ, в то время как выходцы из Google, работающие в Anthropic, увидели возможность и выкупили мощности. Google осознала ошибку только в четвертом квартале, когда выручка от Gemini достигла \$5 млрд ARR, после чего начала панически скупать землю и турбины для дата-центров.

Фактчек: Google продала Anthropic 1 млн TPU v7; Tesla делает чипы Optimus у Samsung в Texace — □: Октябрь 2025 — Anthropic получает до 1 млн TPU и >1 ГВт в 2026. Контракт Tesla-Samsung \$16,5 млрд на AI6, фабрика Taylor — июль 2025.

ASML и фундаментальный предел производства чипов

Смещение бутылочных горлышек 34:37: Ограничения масштабирования постоянно меняются: упаковка CoWoS -> электроэнергия -> постройка дата-центров. К 2028-2030 годам абсолютным и непреодолимым пределом станет производство EUV-литографов компанией ASML.

Математика EUV-литографии для 1 ГВт 37:16: Для создания 1 ГВт дата-центров на чипах Nvidia Rubin требуется 55 000 пластин 3-нм логики, 6 000 пластин 5-нм логики и 170 000 пластин DRAM-памяти. Это требует ≈2 млн проходов через EUV-установки. Для обеспечения 1 ГВт нужно 3,5 EUV-машины. Парадокс в том, что 1 ГВт инфраструктуры стоит 50 млрд, а 3,5 EUV-машины — всего 1,2 млрд.

Прогнозы производства ASML до 2030 года 42:14: Сейчас ASML производит 70 EUV-машин в год, в следующем произведет 80, а к 2030 году выйдет на ≈100 машин в год. Общий мировой парк составит ≈700 установок. Это позволит вводить максимум 200 ГВт ИИ-мощностей ежегодно. План Сэма Альтмана по вводу 52 ГВт в год к концу десятилетия абсолютно реалистичен и займет около 25% мирового производства чипов.

Фактчек: ASML делает 70 EUV-машин в год сейчас, ~100 к 2030, парк ~700 → потолок 200 ГВт/год — △: «70 сейчас» завышено — фактические отгрузки ~44 в 2024 и ~50 в 2025, цель 90/год заявлена на 2027–28; «~100 к 2030» реалистично, но сроки

жестче. Gemini ставит $\square$, хотя её же данные (48 в 2025, 60 в 2026) противоречат «70». Расчёт $700 \div 3,5 = 200$ ГВт — сценарий при отдаче всего парка под ИИ и неизменных техпроцессах, а не физический предел (GPT).

Невозможность быстрого масштабирования ASML 46:43: EUV-литограф — самое сложное устройство, созданное человечеством. Цепочка поставок включает более 10 000 компаний. Линзы Carl Zeiss состоят из идеальных слоев молибдена и рутения, источник света Syner испаряет капли олова лазером, а каретка с кремниевой пластиной движется с ускорением 9G, требуя точности позиционирования до 3 нанометров. Масштабировать такое производство вливанием денег невозможно.

Фактчек: Линзы Zeiss из слоёв молибдена и рутения — $\triangle$: В EUV нет линз — только зеркала; многослойка — ~40-50 пар молибден/кремний, рутений — лишь защитный верхний слой. 9G и позиционирование в единицы нм — верно по порядку величины. Claude и GPT единомышленны, Gemini не проверяла.

Альтернативные пути: старые техпроцессы и Китай

Почему нельзя использовать старые 7-нм фабрики 55:44: Разница между чипами B100 и A100 (7 нм) в чистых FLOPS составляет 3x, но в реальном инференсе (на моделях DeepSeek или Kimi) B100 быстрее в 20 раз. Главное ограничение — пропускная способность памяти и сети. Внутри современного чипа данные передаются на скорости десятков ТБ/с, между чипами — единицы ТБ/с, между стойками — сотни ГБ/с. Разделение модели на множество старых чипов убивает эффективность из-за сетевых задержек.

Китайская полупроводниковая угроза 1:05:42: Сейчас Китай использует DUV-машины ASML для производства 7-нм и 14-нм чипов. К 2030 году Китай полностью локализует производство DUV (≈ 100 машин в год), но массовое производство EUV маловероятно из-за 5-7 летнего этапа «производственного ада».

Сценарии гонки США и Китая 1:11:19: При быстром взлёте ИИ (fast takeoff) США безоговорочно побеждают за счёт вливания \$1 трлн в инфраструктуру и эффекта сложного процента от выручки моделей. Если же создание AGI затянется до 2035 года, Китай может перегнать Запад за счёт полностью вертикально интегрированной, независимой и масштабированной цепочки поставок, не зависящей от европейских и японских подрядчиков.

Кризис памяти (HBM) и удар по потребительской электронике

Невозможность замены HBM на DDR 1:16:21: ИИ-ускорители не могут перейти на обычную память. Стек HBM4 даёт пропускную способность 2,5 ТБ/с на 13 мм периметра чипа. Обычная DDR5 на той же площади даст лишь 64-128 ГБ/с.

Пропускная способность критична для загрузки весов модели и KV-кэша, без нее вычислительные ядра будут простаивать.

Фактчек: HBM4 = 2,5 ТБ/с на стек; HBM требует 4х площади пластины vs DRAM

— Δ : Направление верно, цифры натянуты на 20–30%: JEDEC HBM4 $\approx$ 2 ТБ/с (у Rubin $\sim$ 1,6 ТБ/с на стек), 2,5 — разгонная верхняя граница. По площади Gemini даёт $\square$ со ссылкой на Micron (3:1–4:1), Claude и GPT — $\sim$ 3х и зависимость от поколения и выхода годных; консенсус — «3–4х», не ровно 4.

Уничтожение рынка смартфонов и ПК 1:23:08: В 2026 году 30% всех капитальных затрат Big Tech пойдет на закупку памяти. HBM требует в 4 раза больше площади кремниевой пластины, чем обычная DRAM. Это приведет к кратному росту цен на память. Себестоимость iPhone вырастет на \$150. Глобальные продажи смартфонов рухнут с 1,4 млрд до 500-600 млн штук в год (в основном за счет вымирания бюджетного сегмента).

Фактчек: Смартфоны рухнут с 1,4 млрд до 500-600 млн, iPhone +\$150 — $\square$:

Базовая цифра неверна — мировые поставки $\sim$ 1,2–1,25 млрд/год. Кратное подорожание DRAM — факт, но падение рынка вдвое — экстраполяция одного фактора без учёта удлинения сроков владения, урезания RAM в бюджетниках и б/у. Спекуляция аналитика, не прогноз.

Задержки в строительстве фабрик памяти 1:28:43: Производители памяти (Samsung, SK Hynix, Micron) не закладывали новые фабрики в 2023 году из-за убытков и низких цен. Строительство чистых комнат занимает 2 года, поэтому новые объемы памяти появятся на рынке не раньше конца 2027 года.

Оценка планов Илона Маска по TeraFab 1:32:40: План Маска построить фабрику на 1 млн пластин в месяц упирается в технологии. Он сможет быстро возвести чистые комнаты, но разработка техпроцесса и интеграция оборудования требуют компетенций уровня TSMC, которые невозможно получить с нуля. Надежда индустрии на 2030-е годы — переход на 3D DRAM, что потребует полного переоборудования всех существующих фабрик.

Энергетика: почему электричество не станет проблемой

Скрытые резервы энергосетей 1:42:47: Энергия не станет финальным ограничением. Энергосистема США (около 1 ТВт) имеет 20% резервных мощностей, которые простаивают почти весь год и нужны только для пиковых нагрузок летом. Массовая установка промышленных батарей позволит сгладить пики и отдать эти 20% ($\approx$ 200 ГВт) дата-центрам.

Фактчек: 20% резерва энергосистемы США ($\sim$ 200 ГВт) можно отдать дата-

центрам через батареи — $\square$: Оба проверявших рецензента против. По данным ICF (2026) реальный резерв упал до $\sim$ 3% ($\sim$ 26 ГВт), в PJM и ERCOT свободных мощностей для новых нагрузок нет. Резерв — это мощность для надёжности, а не энергия: 200 ГВт

круглосуточно = ~1750 ТВт·ч/год, ~40% годовой генерации США; батареи переносят энергию во времени, а не создают её.

Локальная генерация (Behind-the-meter) 1:44:58: Половина новых мощностей к концу десятилетия будет генерироваться локально, в обход общих сетей. Будут использоваться авиационные турбины (aeroderivatives), корабельные двигатели, топливные элементы Bloom Energy и поршневые двигатели. Стоимость такой энергии может достигать \$3500 за кВт (в 2 раза дороже газовых турбин комбинированного цикла), но в ТСО кластера это добавит лишь 10 центов к стоимости часа работы GPU, что абсолютно неважно на фоне ценности ИИ-моделей.

Дефицит рабочих рук и модульная сборка 1:50:51: Главная проблема инфраструктуры — нехватка сотен тысяч электриков и строителей. Решением станет модульная сборка дата-центров. Целые ряды серверов (до 1 МВт на стойку для будущих чипов Nvidia Kyber) с интегрированным жидкостным охлаждением, сетью и питанием будут собираться на заводах в Азии и поставляться в США в готовом виде, минимизируя ручной труд на местах.

Дата-центры в космосе и архитектура кластеров

Критика космических дата-центров Маска 1:54:24: Идея выноса вычислений в космос бессмысленна в этом десятилетии. Энергия там бесплатная, но GPU крайне ненадежны (15% чипов Blackwell требуют замены или перезагрузки). Отправка в космос добавит 6 месяцев логистической задержки, что уничтожит 10% полезного жизненного цикла чипа. Космические лазеры для связи спутников дороже и менее надежны, чем земная оптика. Кроме того, в космосе сложнее отводить тепло, тогда как на Земле иммерсионное охлаждение позволяет достичь плотности 2 Вт на квадратный миллиметр чипа.

Топологии кластеров Nvidia и Google 2:06:13: Архитектуры масштабирования (scale-up) сильно различаются. Nvidia NVL72 объединяет 72 GPU по принципу «каждый с каждым» на скорости терабайтов в секунду. Google TPU использует топологию тора: кластер объединяет тысячи чипов, но сигнал идет через 6 соседей с блокировками и задержками.

Преимущество компактных моделей в RL 2:10:42: Лабораториям выгоднее обучать меньшие модели (например, GPT-5.4 вместо огромных монолитов). Меньшая модель требует в 2 раза больше прогонов (rollouts) при обучении с подкреплением (RL), но завершает их быстрее. Это позволяет быстрее внедрить модель в разработку следующих поколений, радикально ускоряя цикл ИИ-исследований.

[2:14:30-2:18:18] смолток о клиентах SemiAnalysis, хедж-фондах и шутки про Леопольда Ашенбреннера.

Будущее TSMC, роботы и геополитика

Apple теряет приоритет у TSMC 2:18:53: Исторически Apple первой получала новые техпроцессы. В поколении 2 нм (A16) ИИ-чипы впервые заберут большую часть объемов. Доля Apple в выручке TSMC будет падать, и компания лишится привилегий. Если бы Huawei не попала под санкции в 2019 году, она бы обошла Apple и стала крупнейшим клиентом TSMC, так как обладает лучшими в мире (после Nvidia) ИИ-исследователями и сетевыми технологиями.

Фактчек: Apple вытесняется на 2 нм; в поколении «2 нм (A16)» ИИ-чипы заберут объёмы — △: N2 и A16 — разные техпроцессы (A16 — класс 1,6 нм с обратной подачей питания), в пересказе они смешаны. Переход Apple на новое поколение не доказывает «вытеснение»; данных о контрактах нет. Проверил только GPT.

Централизованный интеллект для роботов 2:24:18: Интеллект миллионов будущих человекоподобных роботов будет централизован. Тяжелые модели в облаке будут выполнять сложное планирование с высокой эффективностью (batching), отправляя команды локальным чипам робота 10 раз в секунду. Илон Маск производит чипы для своих роботов на заводах Samsung в Техасе, чтобы не конкурировать за мощности TSMC и диверсифицировать геополитические риски.

Риски блокады Тайваня 2:28:41: Идея эвакуации инженеров TSMC в случае войны не спасет мировую экономику. Оборудование ASML требует микросхем, которые производятся только на Тайване (змея, кусающая себя за хвост). Потеря острова обрушит мировой ВВП и снизит ввод новых ИИ-мощностей с сотен гигаватт до ничтожных 10-20 ГВт в год на базе фабрик Intel и Samsung.

ОБЩИЕ ПРОВЕРКИ

Gemini \$5 млрд ARR; 15% чипов Blackwell требуют замены/ребута; H100 при AGI окупится за месяцы — □: Публично не проверяется — инсайд аналитика и условный сценарий. Окупаемость через замену белых воротничков — гипотеза, не расчёт.

ЛИЧНО ДЛЯ ТЕБЯ

Память и техника: покупать, если реально нужно, — но не впрок: Кратное подорожание DRAM/NAND — факт, а новые фабрики памяти физически не дадут объёмов до конца 2027. Если планируешь Mac/ПК на 64–128 ГБ RAM, NAS, телефон или бытовую электронику в новый дом — откладывать невыгодно (Claude, Gemini). GPT против закупок «на всякий случай» под сценарий «смерти смартфонов»: масштаб подорожания не установлен, ориентируйся на цену нужной конфигурации сейчас.

Выбирай модель по стоимости решённой задачи, а не по цене токена: Дешёвого инференса топ-моделей в 2026–27 не жди — цена за токен скорее

стоит, чем падает. По умолчанию гони самую сильную модель (Opus, GPT-5.x): одна переделка и полчаса твоей валидации дороже разницы в токенах. Но это аргумент про твоё время, а не про Алчиана-Аллена — универсального преимущества дорогой модели нет, на рутине дешёвая может выигрывать.

Дом в Лимасоле: солнце + батарея — не эко-жест, а окупаемость 5-7 лет:

Логика behind-the-meter — ровно кипрский кейс: изолированная дорогая сеть, сильное солнце, тарифная разница день/ночь. Заложил кабельные трассы, место под инвертор и батарею сейчас — после отделки это в разы дороже.

Физический потолок AGI — трезвее, чем в пересказе: EUV-литограф действительно нельзя ускорить деньгами (10 000 поставщиков, оловянная плазма, зеркала с точностью в атомы), и ASML выпускает ~50 машин в год, а не 70 — сроки ещё жёстче. Но «200 ГВт/год» — сценарий, а не закон, а «20% резерва сети» — ошибка: энергия и подключения остаются реальным ограничением. Любой прогноз «быстрого взлёта» дисконтируй на оба фактора.

КОНСЕНСУС РЕЦЕНЗЕНТОВ

Все сошлись: дефицит памяти реален и затянется до 2027, EUV-литография — ключевое горлышко (но цифра «70 машин сейчас» завышена), сделки Anthropic-TPU и Tesla-Samsung подтверждены, «линзы из молибдена/рутения» — ошибка (зеркала, Mo/Si). Разошлись: GPT единственный отверг применение эффекта Алчиана-Аллена (□) и смешение N2/A16; Gemini дала □ по «HBM 4x» и ASML, где Claude/GPT видят завышение на 20–30%; по практике Claude и Gemini советуют покупать железо сейчас, GPT — не закупаться под катастрофический сценарий. Тезис про 20% энергорезерва оба проверявших (Gemini, GPT) отвергли независимо и разными аргументами — это самое надёжное опровержение в выпуске.