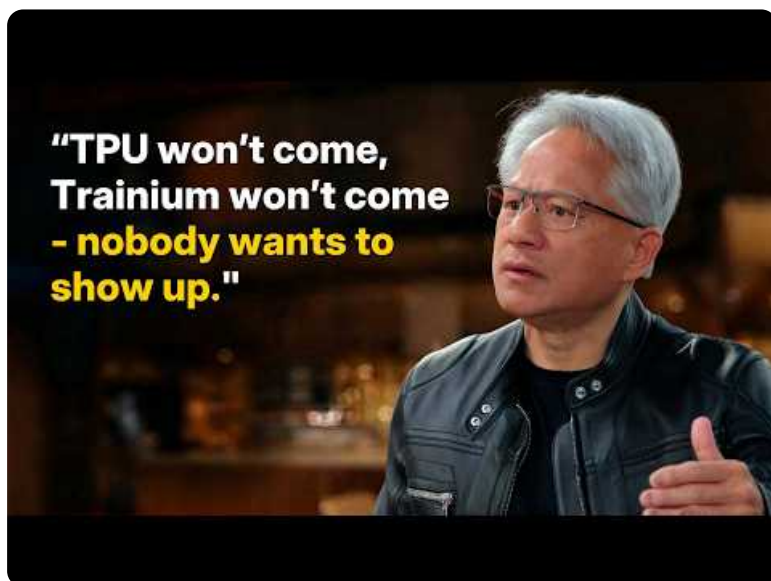


Jensen Huang - Will Nvidia's moat persist?

@DwarkeshPatel · 2026-04-15 · <https://www.youtube.com/watch?v=Hrbq66XqtCo>



Интервью Джансена Хуанга (CEO Nvidia) подкастеру Дваркешу Пателю. Главная тема выпуска — устойчивость доминирования Nvidia на рынке ИИ-вычислений перед лицом конкуренции со стороны кастомных чипов (TPU, ASICs), ограничений цепочек поставок и геополитических барьеров.

Хуанг подробно объясняет механику работы глобальной цепочки поставок от TSMC до дата-центров, доказывает превосходство гибкой архитектуры CUDA над узкоспециализированными решениями и вступает в жесткий спор с ведущим о целесообразности и последствиях экспортных ограничений чипов в Китай.

Роль Nvidia и коммодитизация софта

Трансформация электронов в токены не поддается коммодитизации

00:00: Процесс превращения электричества в ценные ИИ-токены требует колоссальных инженерных и научных усилий. Nvidia выступает связующим звеном в этой цепи, делая ровно столько, сколько необходимо для обеспечения невероятных вычислительных возможностей, и делегируя остальное партнерам. Экосистема компании охватывает все пять уровней ИИ-индустрии: от поставщиков компонентов до разработчиков моделей и приложений.

ИИкратно увеличит рынок корпоративных софтверных инструментов

03:13: Большинство современных софтверных компаний (Excel, Cadence, Synopsys) создают инструменты для людей. В будущем количество пользователей этих инструментов вырастет экспоненциально за счет ИИ-агентов. Ограничение в виде нехватки инженеров исчезнет, так как агенты будут массово использовать существующий софт для исследования

пространства дизайна. Это приведет к взрывному росту числа лицензий и инстансов корпоративного ПО.

Фактчек: Уход Anthropic на TPU/Trainium — «историческая ошибка» Nvidia — □: Anthropic ушёл к Google и Amazon в первую очередь из-за их инвестиций с привязкой к облакам, а не из-за скупости Nvidia. Самокритичная упаковка удобна, но причинность спорна (Claude).

Управление цепочками поставок и масштабирование

Долгосрочные контракты на сотни миллиардов долларов — это следствие, а не причина рва Nvidia 04:54: Поставщики (TSMC, производители памяти и упаковки) готовы инвестировать в расширение производства исключительно под Nvidia, потому что видят масштаб конечного спроса. Хуанг тратит огромную часть времени на обучение и синхронизацию CEO компаний-партнеров, объясняя им будущие потребности рынка. Конференция GTC служит физической точкой сборки, где поставщики компонентов (upstream) видят реальных создателей ИИ-стартапов (downstream) и убеждаются в реальности прогнозов.

Узкие места в производстве чипов решаются за 2-3 года 08:49: Мгновенный спрос на ИИ-вычисления сейчас превышает мировое предложение, что является позитивным сигналом для индустрии. Проблемы с упаковкой CoWoS, которые долгое время были главным бутылочным горлышком, успешно решены путем многократного удвоения мощностей. TSMC теперь масштабирует технологии упаковки синхронно с производством логики и памяти.

Фактчек: Разница в литографии Hopper→Blackwell «около 75%» — △: Hopper (4N) и Blackwell (4NP) — фактически один узел TSMC; кристалл 104 vs 80 млрд транзисторов (+30%), сборка из двух — +160%. «75%» не является осмысленной метрикой (GPT), но суть — рост за счёт архитектуры и сетей, а не транзисторов — верна у всех трёх.

Префетчинг узких мест меняет структуру рынка 12:13: Nvidia заранее инвестирует в технологии, которые станут критическими через несколько лет. Инвестиции в Lumentum, Coherent и экосистему кремниевой фотоники позволили выстроить новую цепочку поставок вокруг TSMC. Компания лицензирует свои патенты (например, технологию COUPE) партнерам, чтобы поддерживать открытость и готовность экосистемы к будущим масштабам.

Фактчек: Валовая маржа Nvidia ~70% — □: GAAP-маржа последних кварталов 72–75%. Но вывод «маржа оправдана лучшей TCO» — □: гиперскейлеры не публикуют внутренние цифры по TPU/Trainium, а отсутствие конкурента в InferenceMAX не доказывает его проигрыш; «ускорение в 2 раза = удвоение выручки» — упрощение (GPT, Claude).

Настоящим ограничением для ИИ станут энергетика и рабочие руки 13:26: В отличие от производства кремния или EUV-литографов, которые можно

масштабировать за пару лет при наличии спроса, строительство новых дата-центров упирается в базовую инфраструктуру. Главный дефицит будущего — это электрики, сантехники и доступ к электроэнергии. Невозможно реиндустриализировать США и построить фабрики ИИ без решения фундаментальных энергетических проблем, требующих долгих лет на реализацию.

Фактчек: Nvidia лицензирует партнёрам собственную технологию COUPE — □: COUPE — платформа интеграции кремниевой фотоники TSMC, а не Nvidia. Ошибку отметил только GPT; Gemini и Claude не проверяли — вероятно, искажение в пересказе.

Архитектура GPU против специализированных TPU/ASIC

Ускоренные вычисления охватывают гораздо больше, чем ИИ 16:39: В отличие от TPU, которые созданы исключительно для тензорных операций, архитектура Nvidia универсальна. Она ускоряет молекулярную динамику, квантовую хромодинамику, обработку неструктурированных данных и физику частиц. Эта универсальность позволяет облачным провайдерам сдавать системы Nvidia в аренду клиентам из любых индустрий, гарантируя стопроцентную утилизацию оборудования.

Фактчек: Китай производит 60% массовых чипов мира и имеет половину ИИ-исследователей — △: Доля КНР в мощностях зрелых узлов (28 нм+) ~30–35%, к 2027 прогноз ~40%; 60% — цифра для отдельных сегментов. «Половина исследователей» — ~47% топ-исследователей получили бакалавриат в Китае (MacroPolo), но большинство работает в США.

Программируемость CUDA — ключ к алгоритмическим прорывам 21:01:

Закон Мура дает прирост производительности лишь на ≈25% в год. Прыжок энергоэффективности в 50 раз при переходе от Hopper к Blackwell достигается за счет новых алгоритмов (Mixture of Experts), распараллеливания и глубокого ко-дизайна железа, сетей (NVLink, Spectrum-X) и софта. Если бы архитектура не была гибко программируемой, исследователи не смогли бы изобретать новые механизмы внимания или гибридные модели.

Фактчек: Blackwell в 50х энергоэффективнее Hopper — △: Это системный бенчмарк (GB200/GB300 NVL72, FP4) против 8xH100 (FP8) на подобранных сценариях инференса, а не чип к чипу; реальные ускорения 2–10х. Gemini считает цифру подтверждённой, Claude и GPT — маркетинговой верхней границей.

Фактчек: Китай застрял на 7 нм → в 10х меньше флопсов (Дваркеш) — △: SMIC делает 7 нм и 5-нм-класс на DUV-мультипаттернинге с плохим выходом; разница 7 vs 3 нм — ~2–3х по плотности/энергоэффективности, не 10х.

[23:54-24:27] рекламная интеграция Crusoe.

Инсталл-база делает CUDA незаменимой даже для гиперскейлеров 24:27:

Несмотря на то, что крупные лаборатории (OpenAI, Google) пишут собственные низкоуровневые ядра (Triton), они продолжают использовать Nvidia из-за гигантской установленной базы в сотни миллионов GPU. Разработчики фреймворков хотят, чтобы их код работал везде — от облака до робототехники. Nvidia активно контрибьютит в открытые фреймворки (Triton, vLLM, NeMo RL) и обеспечивает самую надёжную кодовую базу.

Фактчек: Избыток энергии компенсирует отставание в техпроцессе (Хуанг) — □: Масштабирование кластеров старых чипов упирается в HBM (Китай зависит от SK hynix/Samsung, отрезанных экспортным контролем), interconnect, синхронизацию и надёжность; больше чипов ≠ пропорционально больше полезных вычислений. GPT ставит □, Claude — △ с теми же аргументами.

Nvidia обеспечивает лучшую совокупную стоимость владения (TCO) 30:35:

Маржинальность Nvidia в 70% оправдана тем, что компания предоставляет максимальную производительность на доллар и на ватт. Инженеры Nvidia помогают ИИ-лабораториям оптимизировать их стек, что часто ускоряет работу моделей в 2-3 раза, напрямую удваивая выручку клиентов. Ни один конкурент (включая Trainium или TPU) до сих пор не смог публично доказать свое преимущество в стоимости инференса на независимых бенчмарках вроде InferenceMAX.

Фактчек: Продажи ИИ-чипов Huawei бьют рекорды — □: Прогноз роста выручки от ИИ-чипов на 60% в 2026, до \$12 млрд, на волне Ascend 950PR и DeepSeek V4, оптимизированной под железо Huawei (Gemini).

Инвестиции, облака и распределение дефицитных GPU

Использование TPU компанией Anthropic — это следствие исторической ошибки Nvidia 35:25: На раннем этапе Anthropic требовались миллиарды долларов в виде вычислительных мощностей, которые могли предоставить только Google и AWS (в обмен на использование их чипов). В прошлом Nvidia не имела финансовой возможности и философии для таких огромных венчурных инвестиций. Теперь компания исправила эту ошибку, вложив 30 млрд в OpenAI и 10 млрд в Anthropic, чтобы поддержать их масштабирование на своей платформе.

Фактчек: Nvidia вложила 30 млрд в OpenAI и 10 млрд в Anthropic — □: Anthropic — до 10 млрд (ноябрь 2025); OpenAI — изначально LOI на «до 100 млрд», в реальном раунде начала 2026 сумма усохла до ~\$30 млрд. Gemini и Claude подтверждают, GPT не смог проверить по своим данным; важно различать объявленные обязательства и фактически переведённые деньги.

Фактчек: Claude Mythos находит zero-day во всех популярных ОС, включая OpenBSD — □ с оговоркой: Mythos Preview (Project Glasswing, апрель 2026) нашёл

тысячи уязвимостей, включая 27-летний баг в OpenBSD. «Во всех популярных ОС» — преувеличение ведущего; GPT не смог подтвердить независимо.

Nvidia принципиально не становится облачным провайдером 43:39:

Философия компании — «делать столько, сколько необходимо, и как можно меньше сверх того». Nvidia финансирует нео-облака (CoreWeave, Nscale, Nebius), потому что без этой поддержки они бы не выжили. Цель компании — создать максимально широкую экосистему и позволить партнерам заниматься финансовым и операционным бизнесом, не конкурируя с ними напрямую.

Фактчек: Интеграция Groq в экосистему Nvidia — □: Декабрь 2025 — неэксклюзивная лицензия плюс переход ключевых инженеров за ~\$20 млрд; в 2026 — Groq 3 LPX в платформе Vera Rubin. Ниша «дорогих быстрых токенов» реальна: Cerebras/Groq продают латентность как отдельный продукт.

Отказ от выбора победителей — урок из эпохи 3D-графики 47:29: На заре развития графических ускорителей существовало 60 компаний, и архитектура самой Nvidia изначально была абсолютно ошибочной. Выживание компании научило Хуанга смирению: невозможно предсказать, кто именно победит на рынке. Поэтому Nvidia инвестирует во все перспективные ИИ-стартапы одновременно.

[50:46-51:19] рекламная интеграция Cursor.

Распределение дефицитных GPU идет по очереди, а не через аукцион

51:19: Nvidia отказывается продавать чипы тем, кто предложит большую цену, считая динамическое ценообразование плохой бизнес-практикой. Аллокация зависит от двух факторов: времени размещения официального заказа (PO) и физической готовности дата-центра клиента принять оборудование.

Предсказуемость релизов позволяет рынку планировать стройки 55:14:

Клиенты точно знают график выхода новых поколений (Vera Rubin в этом году, затем Vera Rubin Ultra, затем Feynman). Это позволяет компаниям уверенно проектировать ИИ-фабрики стоимостью от 10 млн до 100 млрд, зная, что стоимость токена будет гарантированно падать на порядок с каждым годом.

Фактчек: Стоимость токена гарантированно падает на порядок каждый год — □: Тренд реальный, но это не закон природы и не гарантия снижения твоего счёта за API — цены провайдеров зависят от спроса и маржи, а не только от железа (GPT).

Геополитика, экспортные ограничения и Китай (Спор ведущего и гостя)

Угроза кибероружия требует ограничения вычислительных мощностей

57:29: В споре об экспортных ограничениях Дваркеш приводит в пример модель Claude Mythos от Anthropic. Она способна находить уязвимости нулевого дня во всех популярных ОС, включая защищенную OpenBSD. Доступ Китая к передовым

чипам для обучения подобных моделей и массового запуска ИИ-хакеров создает прямую угрозу кибербезопасности США.

Китай обладает достаточными ресурсами для развития ИИ даже под санкциями

59:08: Ответный аргумент Хуанга строится на том, что у КНР уже есть всё необходимое. Страна производит 60% массовых чипов в мире, обладает избытком дешевой электроэнергии и половиной всех ИИ-исследователей планеты. Изоляция китайских специалистов и отказ от научного диалога не сделает мир безопаснее.

Открытый исходный код — лучшая защита от ИИ-угроз **1:00:16:** Идея о том, что ИИ-агенты будут бесконтрольно взламывать системы, игнорирует развитие защитной экосистемы. В будущем один мощный ИИ будет окружен тысячами ИИ-агентов, обеспечивающих кибербезопасность. Для развития этой защиты критически важны open-source модели, крупнейшим контрибьютором которых является Китай.

Нехватка передовой литографии компенсируется избытком энергии

1:04:00: Дваркеш указывает, что из-за отсутствия EUV-литографов Китай застрял на 7 нм, что дает им в 10 раз меньше флопсов, чем у США. Хуанг парирует: ИИ — это параллельная вычислительная задача. Имея пустующие дата-центры и неограниченную энергию, Китай может просто объединить в кластеры большее количество старых чипов (уровня Norper), полностью нивелируя отставание в техпроцессе.

Отказ от китайского рынка угрожает технологическому лидерству США

1:08:25: Уход американских компаний заставляет Китай создавать собственную экосистему (продажи ИИ-чипов Huawei бьют исторические рекорды). Если будущие глобальные ИИ-модели будут оптимизированы под китайское железо, а не под американский стек (CUDA), США потеряют контроль над мировыми стандартами при экспорте технологий на Глобальный Юг. Уступка второго по величине рынка в мире — это стратегическая ошибка, аналогичная потере Америкой рынка телекоммуникаций.

Алгоритмические прорывы важнее «голового» железа **1:10:06:** Большинство достижений в ИИ (например, архитектура DeepSeek) связано с математикой и программированием, а не только с транзисторами. Китайские исследователи, ограниченные в вычислительных мощностях, вынуждены придумывать крайне эффективные алгоритмы.

[1:14:27-1:14:57] рекламная интеграция Jane Street.

Аналогия ИИ с ядерным оружием ошибочна и вредна **1:17:34:** Дваркеш сравнивает экспорт вычислительных мощностей с продажей обогащенного урана. Хуанг называет эту аналогию нелогичной: чипы — это базовый инфраструктурный элемент, как электричество. Демонизация ИИ (страх потери

работы программистами или радиологами) наносит ущерб экономике США, так как ИИ автоматизирует задачи, а не профессии.

Будущее архитектур и жизнь без нейросетей

Разница между поколениями чипов кроется не в транзисторах 1:31:54:

Физическая разница в литографии между Hopper и Blackwell составляет всего около 75%. Заявленный прирост производительности в 50 раз достигается исключительно за счет системной архитектуры, сетей и компьютерных наук.

Nvidia не будет возвращаться к старым техпроцессам ради объема

1:35:27: Несмотря на дефицит передовых мощностей TSMC (N3, N2), Nvidia не станет выпускать новые чипы на старых узлах (например, 7 нм). Разработка каждого поколения требует колоссальных затрат на R&D, упаковку и системную архитектуру, поэтому компания фокусирует все ресурсы только на движении вперед.

Отказ от альтернативных архитектур математически обоснован 1:37:07:

Nvidia обладает ресурсами для параллельной разработки чипов размером с пластину (как у Cerebras) или огромных сборок (как Dojo от Tesla). Однако внутренние симуляции доказывают, что эти подходы работают хуже текущей архитектуры, поэтому компания не распыляет усилия.

Сегментация рынка инференса приведет к появлению сверхдорогих токенов 1:38:11:

Nvidia интегрирует решения вроде Groq в свою экосистему, так как рынок ИИ начал расслаиваться. Появился спрос на токены с мгновенным временем отклика (например, для ИИ-программистов). Компании готовы платить высокую цену (ASP) за скорость, даже если общая пропускная способность системы ниже. Это открывает новую нишу премиального инференса.

Без глубокого обучения Nvidia все равно осталась бы гигантом 1:39:48:

Если бы революция нейросетей не случилась, компания продолжила бы развивать ускоренные вычисления. Закон Мура для CPU мертв, и единственный способ наращивать вычислительные мощности для науки (молекулярная динамика, сейсмология, квантовая химия) — это предметно-ориентированное ускорение через GPU. ИИ стал лишь одним из множества алгоритмов, которые выиграли от демократизации суперкомпьютеров.

ЛИЧНО ДЛЯ ТЕБЯ

Латентность — отдельный товар, но проверяй полное время задачи: Для интерактивной работы с кодом (Claude Code, Codex) быстрый отклик часто важнее «умности на 5%». Попробуй Groq/Cerebras для рутинных итераций, сильную медленную модель оставь для архитектуры. Оговорка GPT: агент теряет время на инструменты и тесты, а не на генерацию — доплачивай за скорость только после замера времени до рабочего результата.

Не экономь на инференсе сейчас: Если задача решается только дорогой reasoning-моделью с кучей итераций — покупай результат сегодня, месяцы ожидания не вернутся. Считай стоимость принятого результата (время до решения, число исправлений, минуты твоей проверки), а не цену токена. Но не закладывайся на «10x дешевле каждый год» как на гарантию.

Спор Хуанг-Дваркеш (с 57:29) — учебник рефрейминга: Дваркеш идёт от аналогии (уран) и конкретного примера (Mythos); Хуанг переопределяет категорию («чипы — как электричество») и уходит в экономические последствия, ни разу не отвечая на сам сценарий «ИИ-хакеры под контролем КНР». Полезно распознавать этот приём у собеседников с финансовым интересом в исходе — и у себя.

Энергия становится главным дефицитом: Не чипы, а электрики, сантехники и мегаватты ограничат ИИ на годы. Для стройки в Лимасоле практический вывод: закладывай избыточные мощности по солнцу и накопителям — локальный инференс и энергонезависимость дома дорожают как опция.

КОНСЕНСУС РЕЦЕНЗЕНТОВ

Все трое сошлись: маржа ~70%, сделка с Groq и тезис «рост Blackwell — за счёт системы, а не литографии» верны, тогда как китайские цифры Хуанга (60% чипов, половина исследователей) завышены, а «50x» — подобранный бенчмарк. Разошлись в строгости: Gemini подтверждает 50x и инвестиции без оговорок, Claude добавляет контекст (HBM/interconnect, LOI 100 млрд→30 млрд), GPT не смог верифицировать Mythos и инвестиции и единственный поймал ошибку с COUPE (это технология TSMC) и жёстко отверг «энергия компенсирует техпроцесс».