

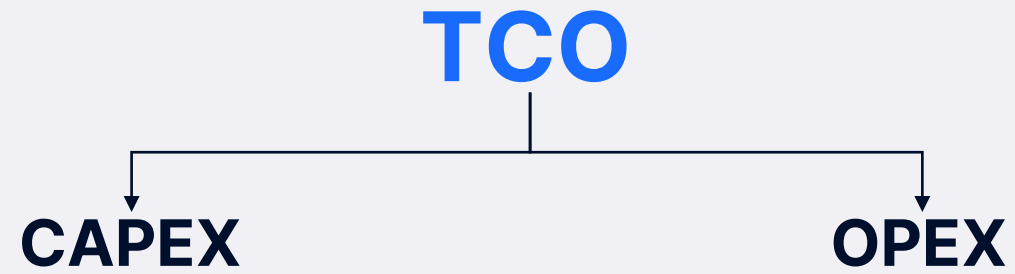
Максим Седых

Руководитель управления инфраструктуры

Железная нехватка

Стратегии перехода от паники к оптимизации

TCO (Total Cost of Ownership*) инфраструктуры облачных сервисов



- Аппаратное обеспечение (доля до 50% от CAPEX)
- Инфраструктура ЦОД
- Софт (лицензии)
- Команда
- Проект, запуск
- Disaster recovery

- Услуги Colocation/ЦОД
- Услуги связи
- ФОТ
- Техническое обслуживание/ремонт
- Поддержка сервисов

*

Совокупная стоимость владения



Рост стоимости серверного оборудования в России

Категория	Рост цен за 2 квартал [*]
Серверная оперативная память DDR4	200-300%
Серверная оперативная память DDR5	400-500%
Твердотельные накопители SSD	30-100%
Флеш-память NAND	до 130%
Серверы	40-60%

Вендоры начали включать в контракты положения, позволяющие им в одностороннем порядке менять цены и отменять заказы в случае дефицита компонентов или непредвиденного роста производственных затрат

Срок действия коммерческих предложений на поставку серверного оборудования снизился в среднем до недели из-за волатильности цен

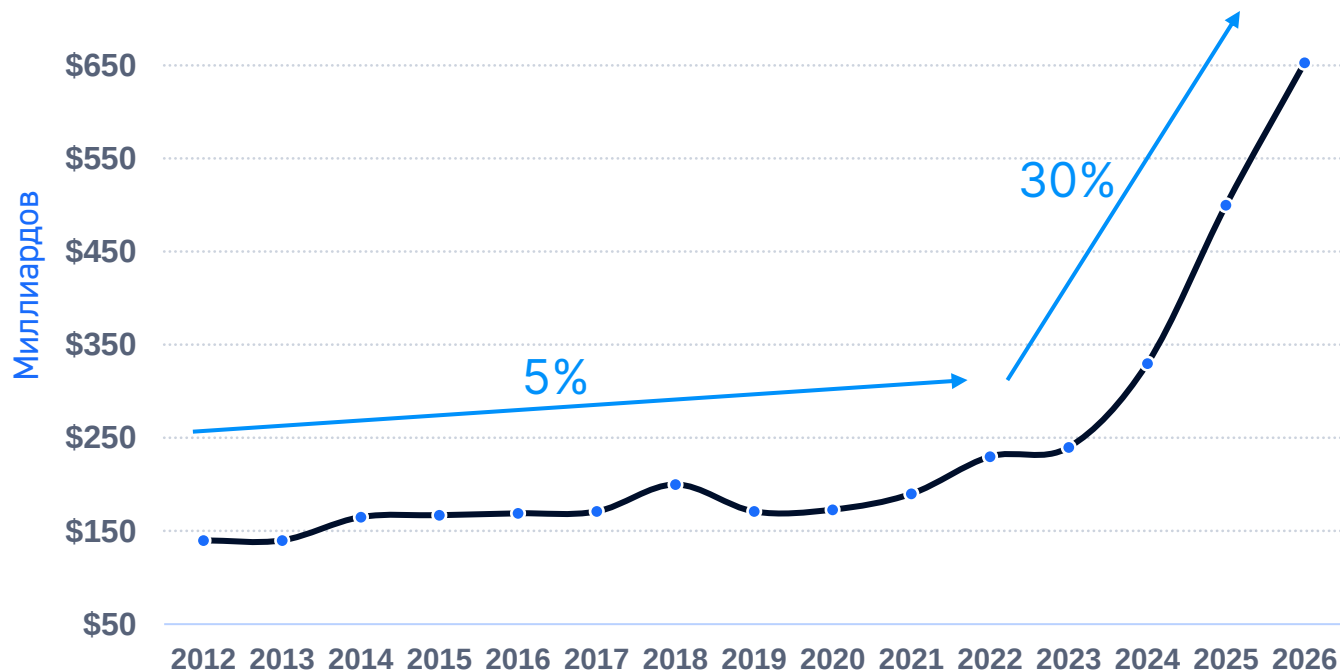
Экспертами прогнозируется дальнейший рост стоимости

^{*} Источник: РБК, 2026



Чем обусловлен рост

Инвестиции в проекты дата-центров *



1 Гонка в секторе ИИ:
совокупный годовой темп роста (CAGR)
инвестиций значительно вырос за последние
годы

2 Память на вес золота:
производители чипов переориентировали
мощности на выпуск высокоскоростной памяти
HBM для ИИ-ускорителей Nvidia и AMD в ущерб
выпуску традиционной памяти DRAM.
HBM полностью законтрактована до 2027 года

3 Проблемы в цепочках поставок:
производство не успевает за спросом,
экспортные ограничения, геополитика,
рост цен на топливо



Источники: Gartner, Inc., 2026
TheNextPlatform, 2025



Влияние на бизнес и прогнозы экспертов

Влияние

Подорожание комплектующих бьёт по всем отраслям в РФ, первыми попадают под удар секторы, обязанные в ближайшее время перейти на российское «железо» — госсектор и субъекты КИИ (согласно Постановлению Правительства РФ №1912)

1

Облачные провайдеры пока сглаживают ценовой шок за счёт закупленного ранее оборудования. Однако стоимость оборудования вырослакратно, тогда как облачные сервисы индексируются значительно плавнее — в среднем около 10% в год

2

Ключевым фактором становится не номинальная цена ресурсов, а предсказуемость затрат и более низкая стоимость владения

3

Прогнозы

Генеральный директор Micron Санджей Мехротра (Sanjay Mehrotra) подчёркивает, что нынешние объёмы выпуска позволяют покрывать лишь от 50 до 67% спроса.

4

По данным IDC* производство DRAM в 2026 году вырастет всего на 16%, что гораздо ниже запроса рынка

Однако спотовые цены на DRAM в течение последних 6 недель вышли на плато и даже снизились благодаря новости о технологии TurboQuant, которую представила компания Google в конце марта. Однако цены на память V-NAND (для производства дисков SSD и NVMe) продолжают стремительно расти

5

Эксперты сходятся во мнении, что рост стоимости серверного оборудования и комплектующих продолжится во втором квартале 2026. Ожидается, что восходящая динамика замедлится во второй половине 2026 года, однако коррекция и нормализация цен прогнозируется не ранее первой половины 2027 года

6

*

Источники: РБК, 2026
Forbes, 2026

*

International Data Corporation ведущая международная аналитическая и консалтинговая компания

Что поможет стабилизировать ситуацию

- **Ввод новых производственных мощностей**

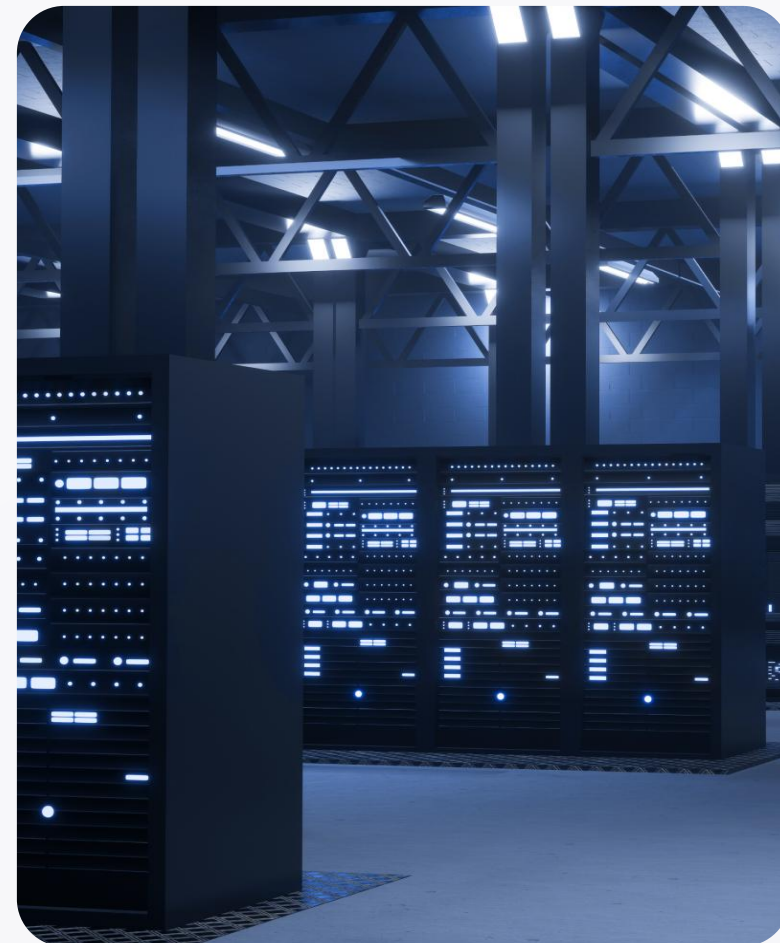
Компании активно инвестируют в строительство новых заводов, однако запуск производства занимает несколько лет. В 2026–2027 годах запланированы запуски производственных площадок, однако большая часть мощностей уже полностью законтрактована. Полноценный эффект на цены ожидается не ранее 2028 года

- **Алгоритм TurboQuant от Google**

Алгоритм экстремального сжатия KV-кэша в больших языковых моделях (LLM) сжимает данные с 16 до 3 бит. Это позволяет в 5–6 раз сократить требования к оперативной памяти (DRAM) и ускорить работу ИИ до 8 раз без потери точности ответов. Полноценный релиз ожидается во второй половине 2026

- **Новый стандарт памяти SPHBM4 от JEDEC**

Стандарт высокоскоростной памяти, обеспечивающий пропускную способность уровня HBM4 при значительно меньшей стоимости и сложности. Призван упростить производство и снизить цену для ИИ-ускорителей



Как бизнесу адаптироваться к новым условиям

Провайдерам услуг

- **Оптимизация существующей инфраструктуры**

Аудит текущей инфраструктуры нередко выявляет неиспользуемые ресурсы, что поможет повысить утилизацию имеющегося оборудования. Внедрение мониторинга ресурсов и FinOps-практик поможет оптимизировать затраты на инфраструктуру

- **Использование предыдущих поколений серверного оборудования**

Память DDR4 дешевле DDR5 в 2-2.5 раза

- **Использование ВТО и OEM сборок, а также REF**

Поможет сэкономить на закупке оборудования, однако вместе с ценой снижается и качество комплектующих

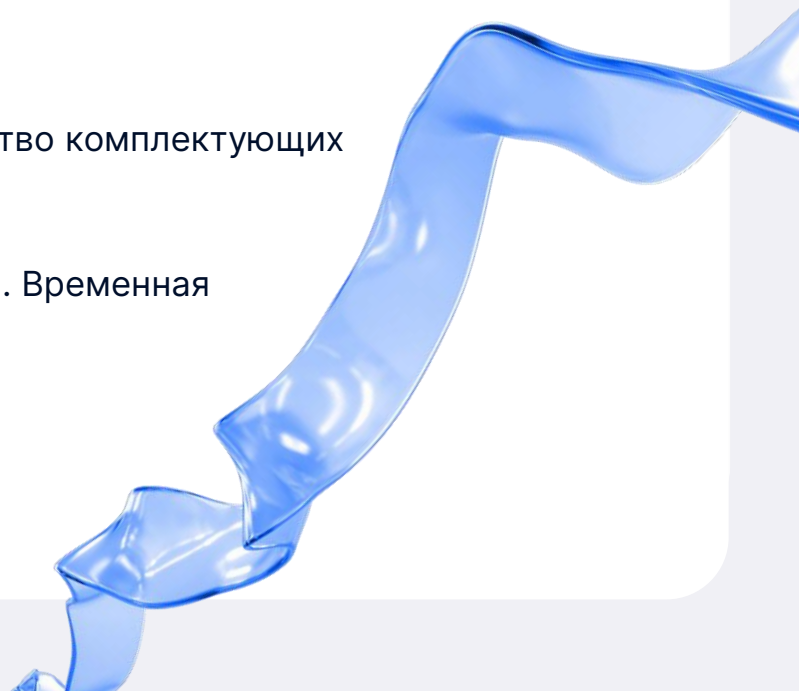
- **Гибридные модели облаков**

Рост цен на оборудование значительно обгоняет индексацию тарифов на облачные услуги. Временная миграция некритичных нагрузок в стороннее облако позволит снизить CAPEX

Потребителям услуг

- **Зафиксировать текущие цены как можно раньше**

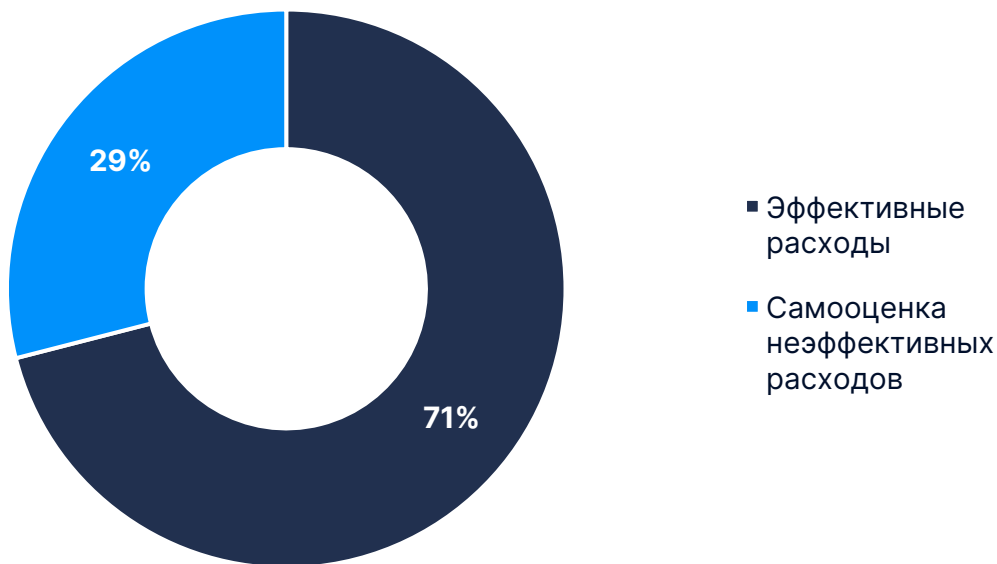
Долгосрочные контракты на аренду серверов, подписанные сегодня, защитят от роста стоимости в 2026–2027 годах



FinOps, ITFM и ITAM

Зачем внедрять FinOps

Эффективность трат на облачные сервисы (IaaS и PaaS) по собственной оценке компаний



- Эффективные расходы
- Самооценка неэффективных расходов

*

Источник: Flexera, 2026

В цифрах

- 84% организаций назвали управление облачными расходами проблемой, а бюджеты превышают плановые значения в среднем на 17%
- 67% организаций тратят на облачные ресурсы больше, чем рассчитывали
- Для 62% опрошенных именно оптимизация расходов на облако является главной задачей в разрезе повышения эффективности трат

15-30%

Среднее снижение расходов на облако от внедрения FinOps 

*

Источник: McKinsey, 2023

Тенденции развития индустрии

Доминирование ИИ

Оборудование оптимизируется для задач ИИ и машинного обучения, требующих высокой производительности, что увеличивает спрос на мощные GPU-серверы

Периферийные вычисления и гравитация данных

Перенос вычислительных мощностей ближе к источнику данных для снижения задержек. Децентрализация вычислений

Жидкостное охлаждение и энергоэффективность

Внедрение систем жидкостного охлаждения и технологий повышения эффективности использования энергии (PUE) становится критическим

Архитектура ARM

Укрепление позиций ARM-процессоров, предлагающих лучшее соотношение производительности на ватт по сравнению с традиционными x86

2,4%

Доля ЦОД и майнинга в РФ от общего энергопотребления. 4,3 ГВт в 2025, рост x2 к 2030

*

*

Источники: «Системный оператор единой энергетической системы» на форуме «Энергетика будущего: вызовы и возможности», 2026

Спасибо!

Буду рад ответить
на ваши вопросы



**Максим
Седых**

Руководитель управления
инфраструктуры

Телефон

+7 (926) 921-14-27

Почта

maxim.sedykh@edgecenter.ru

