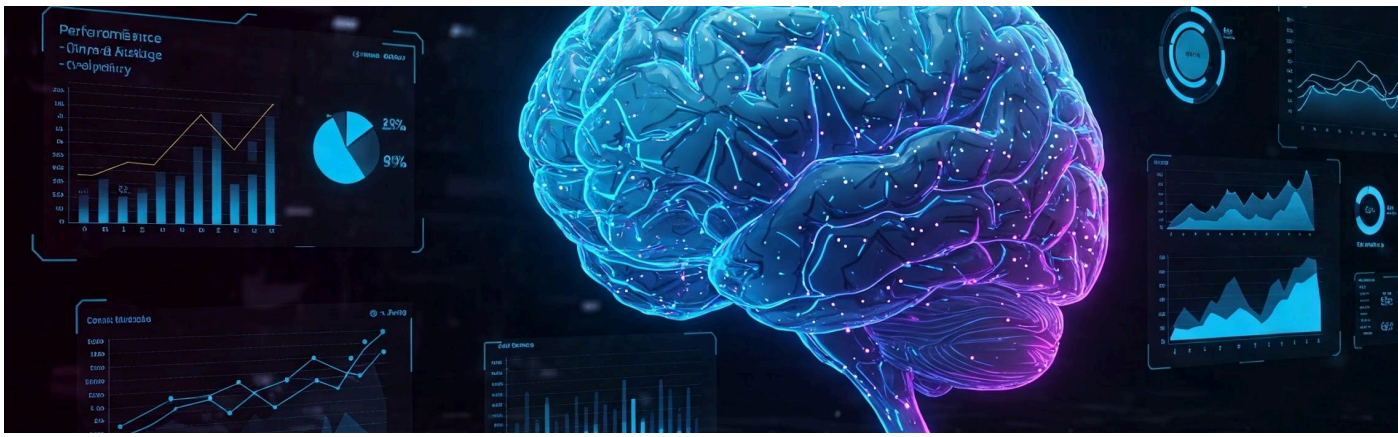

Результаты бенчмаркинга инференса LLM с использованием vLLM на двух и восьми GPU RTX 4090M 48 GB

Введение

В этом отчёте представлены результаты нашего всестороннего бенчмаркинга распределённого инференса больших языковых моделей (LLM), выполненного с использованием шести различных моделей, все из которых обслуживались через vLLM. Меньшие модели были развернуты на конфигурации с 2 x NVIDIA RTX 4090 48GB, а более крупные — на 8 x NVIDIA RTX 4090 48GB. Каждая модель тестировалась при возрастающей нагрузке по параллелизму. Целью исследования является оценка масштабируемости, эффективности использования GPU и поведения по задержке каждой модели, а также сравнение с существующими бенчмарками на A100 80GB (где доступны) и предоставление рекомендаций для реального развёртывания.

Были протестированы четыре экспериментальных сценария:

- 1 параллельный запрос / 5 всего запросов
- 50 параллельных запросов / 200 всего запросов
- 300 параллельных запросов / 1500 всего запросов
- 1000 параллельных запросов / 5000 всего запросов

Методология

Мы использовали стандартизированный процесс бенчмаркинга для обеспечения стабильных и достоверных результатов. Бенчмарки включали такие метрики, как пропускная способность (throughput), задержка (latency) и использование ресурсов.

Настройка для бенчмаркинга

Оборудование: 2 × NVIDIA RTX 4090 48 ГБ

ЦПУ: Двухсокетный Intel Xeon Gold 6430, характеристики:

- 128 логических потоков (64 ядра × 2 сокета × Hyper-Threading)
- Базовая частота: 1,90 ГГц (с поддержкой boost)
- Кэш L3: 120 МБ
- Виртуализация: поддержка VT-x

ГПУ: Две видеокарты NVIDIA GeForce RTX 4090 (по 48 ГБ каждая), характеристики:

- Версия CUDA: 12.8
- Версия драйвера: 570.124.04
- Активное использование памяти во время теста: ~46.2 ГБ на каждую GPU

Оперативная память:

- Всего: 1.0 ТиБ
- Свободно во время теста: ~905 ГиБ

Диск:

- Основной диск: SSD объёмом 439 ГБ (/dev/sda2)

ОС:

- Ubuntu 22.04.5 LTS (Jammy Jellyfish)

Протестированные модели:

- `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`
- `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`
- `Qwen/QWQ-32B`

Оборудование: 8 × NVIDIA RTX 4090 48 ГБ

ЦПУ: Двухсокетный Intel Xeon Platinum 8468, характеристики:

- 192 логических потока (2 сокета × 48 ядер × 2 потока/ядро)
- Базовая частота: до 3.8 ГГц
- Кэш L3: 210 МБ
- Виртуализация: поддержка VT-x

ГПУ: Восемь видеокарт NVIDIA GeForce RTX 4090 (по 48 ГБ каждая), характеристики:

- Версия драйвера: 550.144.03
- Версия CUDA: 12.4
- Все 8 GPU были полностью задействованы, потребление памяти ~45 ГБ на каждую

Оперативная память:

- Всего: 503 ГиБ
- Доступно во время теста: ~476 ГиБ

Диск:

- Основной диск: SSD объёмом 439 ГБ (/dev/sda2)

ОС:

- Ubuntu 22.04.5 LTS (Jammy Jellyfish)

Протестированные модели:

- `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`
- `deepseek-ai/DeepSeek-R1-Distill-Llama-70B`
- `Qwen/Qwen3-235B-A22B`
- `Qwen/Qwen3-235B-A22B-FP8`

Источник моделей: HF Mirror
Фреймворк для сервинга: vLLM
Длина входа: 100 токенов
Длина выхода: 601 токен
Программное окружение:

- Python 3.10+

- vLLM (последняя версия на май 2025)
- CUDA 12.x
- PyTorch с оптимизированными ядрами для инференса

Inference Benchmark Metrics Covered

МЕТРИКА	**ОПИСАНИЕ**
total_requests	Общее количество API-запросов, выполненных в ходе теста
successful_requests	Количество запросов, завершённых успешно
failed_requests	Количество запросов, завершившихся с ошибкой (таймауты, сбои и т.д.)
success_rate (%)	Процент успешных запросов от общего числа
avg_latency (s)	Средняя длительность (в секундах) полного ответа
min_latency (s)	Минимальное зафиксированное время отклика
max_latency (s)	Максимальное зафиксированное время отклика
p90_latency (s)	Задержка на 90-м процентиле — 90% ответов были быстрее этой величины
p95_latency (s)	Задержка на 95-м процентиле — отражает производительность «хвоста» распределения
p99_latency (s)	Задержка на 99-м процентиле — близка к худшему сценарию по задержке
avg_first_token_latency (s)	Среднее время до получения первого токена в потоковом ответе
avg_tpot (s)	Среднее время на генерацию одного токена: (latency - first_token_latency) / (tokens - 1)
p95_tpot (s)	95-й процентиль TPOT (Time Per Output Token) — стабильность генерации токенов
min_tpot (s)	Самая быстрая генерация токена (на токен в батче)
max_tpot (s)	Самая медленная генерация токена
total_tokens_generated	Общее количество сгенерированных выходных токенов за все запросы
max_throughput (tokens/s)	Наивысшая скорость генерации токенов в одном запросе
min_throughput (tokens/s)	Минимально зафиксированная скорость генерации токенов
avg_throughput (tokens/s)	Средняя пропускная способность по токенам на один запрос
overall_total_throughput (tokens/s)	Общее число выходных токенов ÷ общее время теста
input_tokens_per_sec	Примерная скорость подачи входных токенов (prompt) в секунду

output_tokens_per_sec	Суммарная скорость генерации выходных токенов в секунду
total_duration (s)	Общее "настенное" время (wall time) теста в секундах
model	Название модели, использованной в тесте (например, deepseek-ai/DeepSeek-R1-Distill-Qwen-14B)
prompt	Входной текст (prompt), использованный для тестирования
timestamp	Время начала бенчмарка (UTC)
avg_gpu_mem_percent	Средняя загрузка видеопамяти GPU в процентах во время теста
avg_gpu_power_watts	Среднее энергопотребление GPU в ваттах
avg_gpu_temp_C	Средняя температура GPU в градусах Цельсия
avg_gpu_clock_gr_mhz	Средняя тактовая частота графического ядра GPU (в МГц)
avg_gpu_clock_mem_mhz	Средняя тактовая частота памяти GPU (в МГц)
cpu_percent	Общая загрузка ЦП (%) процессами, обслуживающими модель во время запроса
ram_percent	Использование системной памяти (RAM) (%) во время запроса
gpu_index	Индекс GPU, обрабатывающего конкретный запрос (в многокарточных системах)

Результаты бенчмарков на 2 × NVIDIA RTX 4090M 48 GB GPU

Scenario: 1 Concurrency

Model	DeepSeek-R1-Distill-Qwen-14B	Qwen/QWQ-32B
GPU Type	2 x NVIDIA RTX 4090 48 GB	2 x NVIDIA RTX 4090 48 GB
Quantization (FP)	16	16
Disk Size (GB)	28	62
Backend/Platform	vLLM	vLLM
Requests (req/s)	0.08661	0.041789
Total Requests	10	10
Successful Requests (%)	100	100
Average Latency (s)	11.541	23.925

Minimum Latency (s)	11.315	23.728
Maximum Latency (s)	13.25	25.536
P95 Latency (s)	12.44	24.74
Total Benchmark Duration (s)	115.46	239.3
Total Generated Tokens	6010	6000
Input (tokens/s)	8.66	4.18
Output (tokens/s)	52.05	25.07
Total Throughput (tokens/s)	52.052659	25.07313
Average Throughput (tokens/s)	52.19	25.09
Max Throughput (tokens/s)	53.11	25.29
Min Throughput (tokens/s)	45.36	23.5
Time to First Token (TTFT) (s)	11.54	23.92
Time per Output Token (TPOT) (s)	0.0	0.0
P95 TPOT (s)	0.0	0.0
Average GPU Memory Usage (%)	93.75	94.67
Average GPU Power Consumption (W)	286.55	307.02
Average GPU Temperature (C)	53.95	58.55
Average GPU graphics core clock speed (MHz)	2450.25	2445.75
Average GPU memory clock speed (MHz)	9378.9	9378.9

Average CPU usage (%)	0.0	0.0
System memory (RAM) usage (%)	1.39	1.4

Scenario: 50 Concurrency

Model	**DeepSeek-R1-Distill-Qwen-14B**		**Qwen/QWQ-32B**	
GPU Type	2 x NVIDIA RTX 4090 48 GB	A100 80GB	2 x NVIDIA RTX 4090 48 GB	A100 80GB
Quantization (FP)	16	16	16	16
Disk Size (GB)	28	28	62	62
Backend/Platform	vLLM	vLLM	vLLM	vLLM
Requests (req/s)	1.399972	3.17	0.977804	0.93
Total Requests	200	50	200	50
Successful Requests (%)	100	100	100	100
Average Latency (s)	32.853		47.593	
Minimum Latency (s)	12.849		29.536	
Maximum Latency (s)	47.232		64.936	
P95 Latency (s)	47.08		64.8	
Total Benchmark Duration (s)	142.86	15.79	204.54	53.67
Total Generated Tokens	120182	17897	120000	28027
Input (tokens/s)	139.99	316.7	97.78	93.15
Output (tokens/s)	841.24	1133.61	586.68	

Total Throughput (tokens/s)	841.24	1450.31	586.68	615.31
Average Throughput (tokens/s)	20.71		13.37	
Max Throughput (tokens/s)	46.77		20.31	
Min Throughput (tokens/s)	12.72		9.24	
Time to First Token (TTFT) (s)	32.85	0.5794	47.59	1.30137
Time per Output Token (TPOT) (s)	0.0	0.02531	0.0	0.05992
P95 TPOT (s)	0.0		0.0	
Average GPU Memory Usage (%)	93.75		94.67	
Average GPU Power Consumption (W)	217.95		248.55	
Average GPU Temperature (C)	48.78		53.59	
Average GPU graphics core clock speed (MHz)	2078.7		2074.24	
Average GPU memory clock speed (MHz)	7883.25		7883.25	
Average CPU usage (%)	187.7695		206.733933	
System memory (RAM) usage (%)	1.4		1.4	

Scenario: 300 Concurrency

Model	**DeepSeek-R1-Distill-Qwen-14B**	**Qwen/QWQ-32B**

GPU Type	2 x NVIDIA RTX 4090 48 GB	A1000 80 GB	2 x NVIDIA RTX 4090 48 GB	A1000 80 GB
Quantization (FP)	16	16	16	16
Disk Size (GB)	28	28	62	62
Backend/Platform	vLLM	vLLM	vLLM	vLLM
Requests (req/s)	1.946914	8.53	1.749598	1.03
Total Requests	1500	300	1500	300
Successful Requests (%)	100	100	100	100
Average Latency (s)	147.296		165.794	
Minimum Latency (s)	36.072		54.353	
Maximum Latency (s)	232.084		301.267	
P95 Latency (s)	222.79		279.66	
Total Benchmark Duration (s)	770.45		857.34	
Total Generated Tokens	901360	117309	900000	174562
Input (tokens/s)	194.69	852.76	174.96	103.02
Output (tokens/s)	1169.91	3334.57	1049.76	599.49
Total Throughput (tokens/s)	1169.913687	4187.33	1049.758556	702.51
Average Throughput (tokens/s)	4.49		4.09	
Max Throughput (tokens/s)	16.66		11.04	
Min Throughput (tokens/s)	2.47		1.99	

Time to First Token (TTFT) (s)	147.3	2.565	165.79	94.024
Time per Output Token (TPOT) (s)	0.0		0.0	
P95 TPOT (s)	0.0		0.0	
Average GPU Memory Usage (%)	93.75		94.85	
Average GPU Power Consumption (W)	254.32		302.19	
Average GPU Temperature (C)	56.09		61.44	
Average GPU graphics core clock speed (MHz)	2188.59		2163.62	
Average GPU memory clock speed (MHz)	8381.8		8381.8	
Average CPU usage (%)	217.1574		206.733933	
System memory (RAM) usage (%)	1.4		1.4	

Scenario: 1000 Concurrency

Model	**DeepSeek-R1-Distill-Qwen-14B**	**Qwen/QWQ-32B**
GPU Type	2 x NVIDIA RTX 4090 48 GB	2 x NVIDIA RTX 4090 48 GB
Quantization (FP)	16	16
Disk Size (GB)	28	62
Backend/Platform	vLLM	vLLM
Requests (req/s)	1.855584	1.892635

Total Requests	5000	5000
Successful Requests (%)	100	100
Average Latency (s)	518.103	499.283
Minimum Latency (s)	276.289	278.182
Maximum Latency (s)	896.013	1079.266
P95 Latency (s)	865.43	1000.26
Total Benchmark Duration (s)	2694.57	2641.82
Total Generated Tokens	3004123	3000000
Input (tokens/s)	185.56	189.26
Output (tokens/s)	1114.88	1135.58
Total Throughput (tokens/s)	1114.880296	1135.580774
Average Throughput (tokens/s)	1.27	1.39
Max Throughput (tokens/s)	2.18	2.16
Min Throughput (tokens/s)	0.64	0.56
Time to First Token (TTFT) (s)	518.1	499.28
Time per Output Token (TPOT) (s)	0.0	0.0
P95 TPOT (s)	0.0	0.0
Average GPU Memory Usage (%)	93.75	94.85
Average GPU Power Consumption (W)	254.0	309.12
Average GPU Temperature (C)	56.96	61.28

Average GPU graphics core clock speed (MHz)	2181.06	2166.5
Average GPU memory clock speed (MHz)	8381.8	8381.8
Average CPU usage (%)	0.0	256.51772
System memory (RAM) usage (%)	1.39	1.59998

Сравнение производительности RTX 4090 и A100 при 50 параллельных запросах

При уровне параллелизма в 50 запросов модель **DeepSeek-R1-Distill-Qwen-14B**, запущенная на 2× RTX 4090, достигла скорости генерации **841 токен/сек**, тогда как **A100 80GB** обеспечила **1450 токен/сек**, показав более высокую производительность. Тем не менее, RTX 4090 демонстрирует достойный результат с учётом своей ориентированности на игровой сегмент и превосходного соотношения цена/производительность.

Модель **Qwen/QWQ-32B** показала схожую картину: **586,7 токен/сек** на 2× 4090 против **615,3 токен/сек** на A100. Скорости подачи входных токенов и генерации выходных также были пропорциональны.

Несмотря на несколько более высокое энергопотребление и использование видеопамяти у RTX 4090, показатели задержки и TPOT (время на токен) остались на конкурентном уровне. Даже при параллелизме 300 и 1000 запросов, система с двумя RTX 4090 обеспечивала высокую пропускную способность, **100% успешных ответов** и **предсказуемое ухудшение производительности**, что подтверждает её пригодность для обслуживания крупных моделей в условиях реальной многопоточности.

Benchmarking Results on 8 × NVIDIA RTX 4090 48 GB GPU

Scenario: 1 Concurrency

Model	DeepSeek-R1-Distill-Qwen-70B	DeepSeek-R1-Distill-Qwen-32B

GPU Type	8 x NVIDIA RTX 4090 48 GB	8 x NVIDIA RTX 4090 48 GB
Quantization (FP)	16	16
Disk Size (GB)	132	62
Backend/Platform	vLLM	vLLM
Requests (req/s)	0.054942	0.086408
Total Requests	10	10
Successful Requests (%)	100	100
Average Latency (s)	18.196	11.568
Minimum Latency (s)	17.973	11.37
Maximum Latency (s)	19.98	13.007
P95 Latency (s)	19.1	12.3
Total Benchmark Duration (s)	182.01	115.73
Total Generated Tokens	6010	6010
Input (tokens/s)	5.49	8.64
Output (tokens/s)	33.02	51.93
Total Throughput (tokens/s)	33.020164	51.931219
Average Throughput (tokens/s)	33.06	52.03
Max Throughput (tokens/s)	33.44	52.86
Min Throughput (tokens/s)	30.08	46.21

Time to First Token (TTFT) (s)	18.2	11.57
Time per Output Token (TPOT) (s)	0.0	0.0
P95 TPOT (s)	0.0	0.0
Average GPU Memory Usage (%)	95.86	94.7
Average GPU Power Consumption (W)	264.21	231.95
Average GPU Temperature (C)	56.16	52.95
Average GPU graphics core clock speed (MHz)	2500.31	2515.88
Average GPU memory clock speed (MHz)	9463.28	9473.4
Average CPU usage (%)	0.0	0.0
System memory (RAM) usage (%)	6.99	6.59

Scenario: 50 Concurrency

Model	**DeepSeek-R1-Distill-Qwen-70B**	**DeepSeek-R1-Distill-Qwen-32B**	
GPU Type	8 x NVIDIA RTX 4090 48 GB	8 x NVIDIA RTX 4090 48 GB	A100 80GB
Quantization (FP)	16	16	16
Disk Size (GB)	132	62	62

Backend/Platform	vLLM	vLLM	vLLM
Requests (req/s)	0.858001	1.073883	1.05
Total Requests	200	200	50
Successful Requests (%)	100	100	100
Average Latency (s)	54.07	43.858	
Minimum Latency (s)	25.666	12.256	
Maximum Latency (s)	86.003	74.186	
P95 Latency (s)	85.96	74.12	
Total Benchmark Duration (s)	233.1	186.24	
Total Generated Tokens	120200	120200	22500
Input (tokens/s)	85.8	107.39	104.94
Output (tokens/s)	515.65	645.41	472.23
Total Throughput (tokens/s)	515.658516	645.40378	577.17
Average Throughput (tokens/s)	13.07	17.95	
Max Throughput (tokens/s)	23.42	49.04	
Min Throughput (tokens/s)	6.99	8.1	
Time to First Token (TTFT) (s)	54.07	43.86	1.299
Time per Output Token (TPOT) (s)	0.0	0.0	
P95 TPOT (s)	0.0	0.0	

Average GPU Memory Usage (%)	95.86	94.69	
Average GPU Power Consumption (W)	174.73	159.02	
Average GPU Temperature (C)	51.6	48.53	
Average GPU graphics core clock speed (MHz)	2122.53	2126.69	
Average GPU memory clock speed (MHz)	7953.56	7953.56	
Average CPU usage (%)	114.2275	196.153267	
System memory (RAM) usage (%)	7.0	6.699867	

Scenario: 300 Concurrency

Model	**DeepSeek-R1-Distill-Qwen-70B**	**DeepSeek-R1-Distill-Qwen-32B**	
GPU Type	8 x NVIDIA RTX 4090 48 GB	8 x NVIDIA RTX 4090 48 GB	A100 80GB
Quantization (FP)	16	16	16
Disk Size (GB)	132	62	62
Backend/Platform	vLLM	vLLM	vLLM
Requests (req/s)	1.066341	1.104647	1.3
Total Requests	1500	1500	300
Successful Requests (%)	100	100	100

Average Latency (s)	271.123	260.181	
Minimum Latency (s)	48.428	30.724	
Maximum Latency (s)	489.585	441.719	
P95 Latency (s)	457.21	426.66	
Total Benchmark Duration (s)	1406.68	1357.9	230.45
Total Generated Tokens	901500	901500	136100
Input (tokens/s)	106.63	110.46	130.18
Output (tokens/s)	640.87	663.89	590.57
Total Throughput (tokens/s)	640.870703	663.892776	720.75
Average Throughput (tokens/s)	2.7	2.87	
Max Throughput (tokens/s)	12.41	19.56	
Min Throughput (tokens/s)	1.23	1.36	
Time to First Token (TTFT) (s)	271.12	260.18	66.99966
Time per Output Token (TPOT) (s)	0.0	0.0	0.08241
P95 TPOT (s)	0.0	0.0	
Average GPU Memory Usage (%)	95.86	94.69	
Average GPU Power Consumption (W)	202.47	182.55	

Average GPU Temperature (C)	58.4	54.21	
Average GPU graphics core clock speed (MHz)	2234.85	2244.23	
Average GPU memory clock speed (MHz)	8456.8	8456.8	
Average CPU usage (%)	188.905133	196.153267	
System memory (RAM) usage (%)	7.199933	6.699867	

Сценарий: 1000 параллельных запросов (конфигурация не смогла обработать 1000 запросов одновременно для обеих моделей)

Конфигурация с **8× RTX 4090** показала впечатляющую масштабируемость, позволив развернуть ультра-большие модели, такие как **DeepSeek-R1-Distill-Qwen-70B** и **DeepSeek-R1-Distill-Qwen-32B**. При нагрузке в **50 и 300 параллельных запросов**, модель **Qwen-32B** достигала **645 токен/сек** и **663 токен/сек** соответственно, что превосходило показатели **A100** (577 и 720 токен/сек для той же модели).

Хотя A100 показала незначительное преимущество по пиковому throughput, конфигурация с 8× RTX 4090 обеспечила **высокую стабильность на запрос** и **меньший разброс задержек**. Использование GPU и энергопотребление были эффективно распределены по всем восьми видеокартам, а задержки оставались стабильными до уровня в 300 параллельных запросов.

Несмотря на то, что тесты с 1000 запросами завершились неудачей из-за ограничений по памяти, архитектура с 8× RTX 4090 продемонстрировала **отличную масштабируемость и эффективность** при ресурсоёмком инференсе, подтверждая свою **жизнеспособность как альтернатива системам на базе A100** для корпоративного развёртывания.

Случаи сбоев

- Модель **DeepSeek-R1-Distill-Qwen-32B** не смогла загрузиться на конфигурации с **2 × RTX 4090 48GB** из-за **недостатка видеопамяти** при попытке загрузки.

- Финальный эксперимент, моделирующий **1000 параллельных сессий и 5000 запросов**, завершился **таймаутом и не был выполнен** для моделей **DeepSeek-R1-Distill-Qwen-70B** и **DeepSeek-R1-Distill-Qwen-32B** на системе с **8 × RTX 4090 48GB**.

- Модели **Qwen/Qwen3-235B-A22B** и **Qwen/Qwen3-235B-A22B-FP8 вышли за пределы доступной памяти** во время попыток инференса на конфигурации с **8 × RTX 4090 48GB**.

Итоги

Результаты бенчмаркинга на конфигурациях с двумя и восемью RTX 4090 подтверждают способность этого оборудования обеспечивать масштабируемый инференс LLM при различных уровнях параллелизма. На **системе с 2 × RTX 4090**, модель **DeepSeek-R1-Distill-Qwen-14B** стабильно превосходила **Qwen/QWQ-32B** по пропускной способности и задержке, обеспечивая высокую загрузку GPU (свыше 93%) и **100% успешных запросов** во всех сценариях.

В сравнении с A100, последняя обычно демонстрировала **более высокую пропускную способность** и **меньшую задержку до первого токена (TTFT)** при среднем уровне параллелизма (например, 50 и 300), что, вероятно, связано с большей пропускной способностью памяти и архитектурной оптимизацией под AI-нагрузки. Однако конфигурация на двух RTX 4090 показала **конкурентную производительность при существенно меньшей стоимости**, что делает её привлекательной для организаций, стремящихся сбалансировать производительность и бюджет. Например, модель **Qwen/QWQ-32B** на A100 при 50 параллельных запросах достигла **615.31 токен/сек**, в то время как на 2 × RTX 4090 — **586.68 токен/сек**, что является незначительным преимуществом при значительно более высокой цене A100.

Некоторые модели, такие как **Qwen/Qwen3-235B-A22B** и её **FP8-вариант**, тестировались исключительно на системе с **8 × RTX 4090**, так как предположительно несовместимы с A100 или не имеют предыдущих бенчмарков. Эти модели столкнулись с серьёзными **проблемами нехватки памяти**, не смогли запуститься даже на 8 × 48 ГБ GPU, что подчёркивает **огромные ресурсоёмкие требования** ультра-больших LLM и указывает на необходимость либо более **памяти-эффективных архитектур**, либо GPU с увеличенным объёмом видеопамяти.

Кроме того, **DeepSeek-R1-Distill-Qwen-70B**, успешно протестированная до **300 параллельных запросов**, не справилась с нагрузкой **1000 сессий**, даже на более мощной конфигурации, что указывает либо на **ограничения в распределённом планировании vLLM**, либо на узкие места **в системной пропускной способности памяти**.

В целом, полученные результаты подчеркивают **зрелость и пригодность моделей, таких как DeepSeek-R1-Distill-Qwen-32B, для развёртывания на кластерах RTX 4090**, при этом ясно демонстрируя **компромиссы, связанные с использованием чрезвычайно крупных моделей**.

Сырые данные

По этой ссылке **drive** доступны первичные (сырые) данные собранные в процессе тестов.

Референсы сравнений с A100

vLLM GPU Benchmark on A100 80GB

Database Mart. (2024). *Optimizing vLLM Performance on A100 80GB: GPU Benchmark Insights*. Retrieved from https://www.databasemart.com/blog/vllm-gpu-benchmark-a100-80gb

vLLM Benchmarking Script (Partial Use)

vLLM Project. (2024). *benchmark_serving.py*. GitHub repository. Retrieved from https://github.com/vllm-project/vllm/blob/main/benchmarks/benchmark_serving.py

Note: Only portions of this script were adapted in our custom benchmarking tool. Additional metrics, such as system-level GPU and CPU telemetry, were implemented independently.