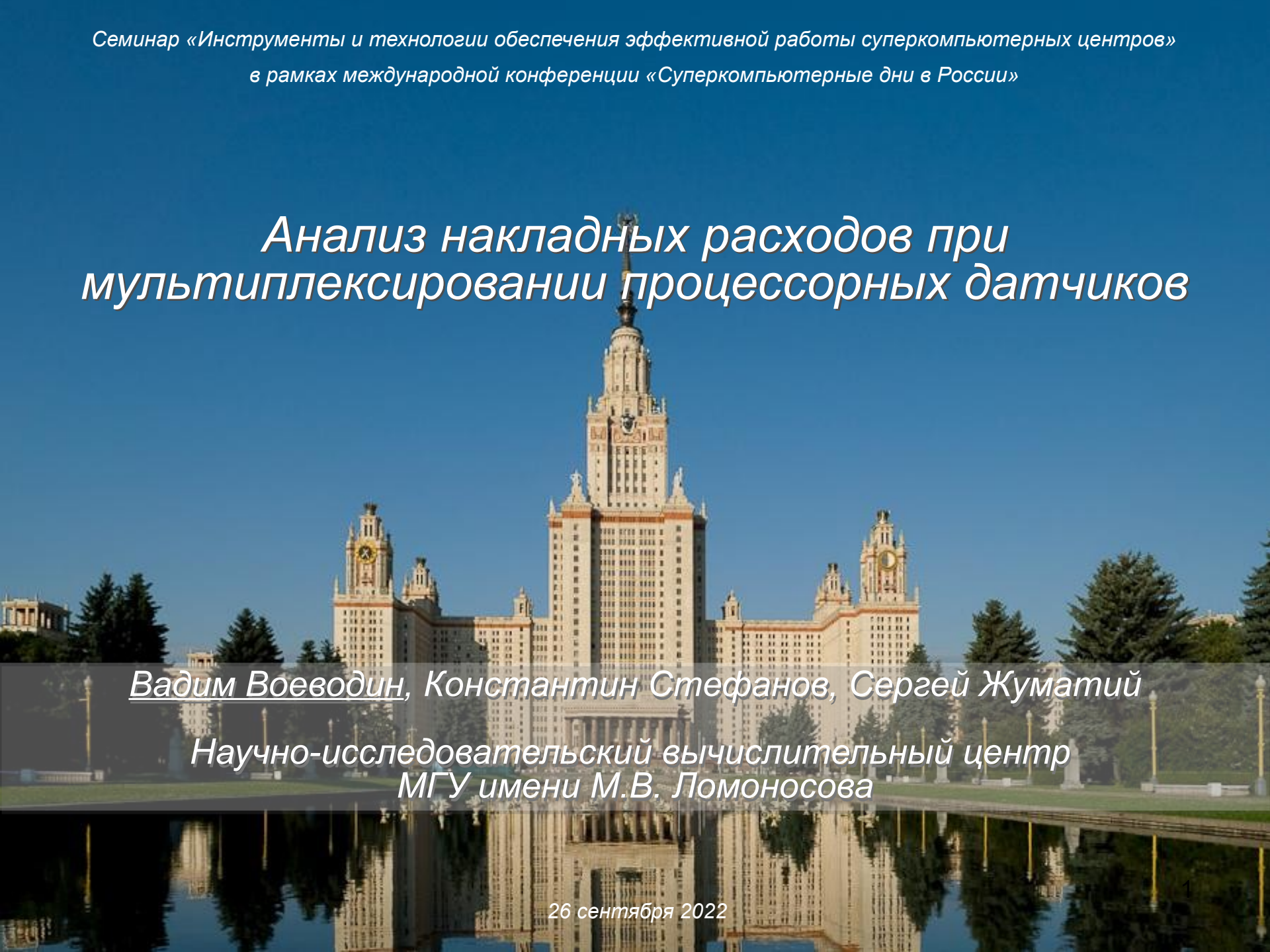
Семинар «Инструменты и технологии обеспечения эффективной работы суперкомпьютерных центров»
в рамках международной конференции «Суперкомпьютерные дни в России»

Анализ накладных расходов при мультиплексировании процессорных датчиков



Вадим Воеводин, Константин Стефанов, Сергей Жуматий

Научно-исследовательский вычислительный центр
МГУ имени М.В. Ломоносова

26 сентября 2022

Для чего нам это нужно?

- Суперкомпьютеры используются далеко не на 100%
- Одна из главных причин – низкая эффективность пользовательских приложений
  - Хотя причины бывают самые разные
- Привычные метрики для оценки использования ресурсов СКЦ не очень удобны для анализа эффективности приложений
  - % занятых узлов ничего не говорит об эффективности приложений
  - Загрузка CPU не позволяет понять, насколько процессор занят полезными вычислениями

- **Нужны новые оценки, которые позволяют быстрее и точнее выявлять проблемы с эффективностью в приложениях**

Вычисление оценок

- Единица измерения – «качество использования» отдельного типа ресурсов отдельно взятой задачей
  - Мы оцениваем, насколько работа с определенным ресурсом мешает полезному использованию CPU/GPU (или насколько полезно задействованы сами CPU/GPU)

- Исследуемые типы ресурсов:
  - CPU
  - Подсистема памяти
  - Коммуникационная сеть (работа с MPI)
  - Ввод/Вывод (I/O)
  - GPU (отдельно процессор и память)

Вычисление оценок

- Единица измерения – «качество использования» отдельного типа ресурсов отдельно взятой задачей
  - Мы оцениваем, насколько работа с определенным ресурсом мешает полезному использованию CPU/GPU (или насколько полезно задействованы сами CPU/GPU)

- Исследуемые типы ресурсов:
  - **CPU**
  - **Подсистема памяти**
  - Коммуникационная сеть (работа с MPI)
  - Ввод/Вывод (I/O)
  - GPU (отдельно процессор и память)

- Оценки CPU и подсистемы памяти – на основе Top-down подхода[1]

$score_{cpu}$ = 100 - Retiring = 100 – 100 * UOPS_RETIRED:RETIRE_SLOTS / (2 * CPU_CLK_UNHALTED:THREAD_ANY)

$score_{mem}$ = Memory bound = 100 * (min(CPU_CLK_UNHALTED:THREAD, CYCLE_ACTIVITY:STALLS_LDM_PENDING) + RESOURCE_STALLS:SB) / CPU_CLK_UNHALTED:THREAD

[1] *https://software.intel.com/en-us/vtune-amplifier-cookbook-top-down-microarchitecture-analysis-method*

Сбор данных

- Для вычисления оценок нужны данные с 5 датчиков (+ 3 доп.)
- ***НО:*** Одновременно можно снимать данные с не более чем 4 процессорных датчиков

Требуется мультиплексирование (чередование наборов датчиков)

- ***НО:*** Мультиплексирование увеличивает накладные расходы. И нет информации, насколько сильно.

Требуется провести анализ накладных расходов при мультиплексировании

Как проводились эксперименты

- Сбор данных – 1 раз в сек. с помощью DiMMon на Ломоносов-2
  - Результаты не зависят от выбранной системы мониторинга, т.к. задержка не там
- Рассматриваем четыре варианта:
  - Без мультиплексирования: PAPI, сбор 4 датчиков (базовый вариант для сравнения)
  - С мультиплексированием:
    - PAPI в автоматическом режиме
    - PAPI в ручном режиме
    - LIKWID в ручном режиме
- Изучаем, насколько замедляется время выполнения набора тестов по сравнению с базовым вариантом

Описание датчика	Именование PAPI	Именование LIKWID
Number of instructions retired	PAPI_TOT_INS	INST_RETIRED_ANY_P
Core cycles when the thread is not in halt state	PAPI_TOT_CYC	CPU_CLOCK_UNHALTED_THREAD_P
L1D data line replacements	PAPI_L1_DCM	L1D_REPLACEMENT
Core-originated cacheable demand requests missed L3	PAPI_L3_TCM	LONGEST_LAT_CACHE_MISS
Retirement slots used	UOPS_RETIRED:RETIRE_SLOTS	UOPS_RETIRED_RETIRE_SLOTS
Core cycles when at least one thread on the physical core is not in halt state	CPU_CLK_UNHALTED:t=1	CPU_CLOCK_UNHALTED_THREAD_P_ANY
Execution stalls due to memory subsystem	CYCLE_ACTIVITY:STALLS_LDM_PENDING	CYCLE_ACTIVITY_STALLS_LDM_PENDING
Cycles stalled due to no store buffers available	RESOURCE_STALLS: SB	RESOURCE_STALLS_SB

Что тестировали

- Анализировали 8 тестов NPB – BT, CG, EP, FT, IS, LU, MG, SP.
  - Реализации на Fortran+MPI
- 11 конфигураций тестов, в зависимости от числа процессов, числа узлов и класса теста.
- Нотация: *<число узлов>-<число процессов на узел>-<класс теста>*
  - 4-1-C – тест на 4 узлах, по 1 процессу на узел, выбран класс C
- Время выполнения – от нескольких секунд до 1.5 часов
  - Чаще всего меньше 10 минут

Где и как тестировали

- Эксперименты проводили на Ломоносов-2, в разделе test/compute
  - Intel Xeon E5-2697 v3, 64 ГБ ОЗУ
  - PAPI 5.6 + LIKWID 5.2.1. CentOS 7, ядро 3.10.0
- Эксперименты повторяли по 10 раз
- Вручную отсеивали аномалии (влияние внешней среды)
- NB: PAPI в автоматическом режиме выдавал некорректные значения
  - Считаем, что проблема не общая, а для конкретных версий OS/PAPI/датчиков

PAPI в автоматическом режиме

Накладные расходы

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	7.19	7.64	1.63	8.55	6.07	7.49	—	-0.61	5.21	6.90	1.47	5.15
CG	4.78	4.66	1.77	4.81	3.21	3.89	6.44	-2.98	-7.02	5.52	4.14	2.66
EP	8.74	4.98	0.82	-0.15	-2.01	6.42	-11.19	-0.55	-8.75	11.31	2.14	1.07
FT	6.66	5.64	1.49	6.74	4.45	5.17	4.65	-1.24	4.84	6.76	—	4.52
IS	7.42	4.46	0.88	6.05	6.36	2.56	6.51	0.00	5.17	5.05	-0.82	3.97
LU	7.36	7.36	1.35	5.82	6.53	4.92	4.69	0.85	2.92	15.38	1.71	5.36
MG	6.50	5.90	2.14	4.75	5.78	7.39	3.63	-4.07	7.18	—	1.88	4.11
SP	5.87	4.78	1.47	6.18	5.93	4.97	—	-3.07	6.31	6.73	0.96	4.02
avg.	6.81	5.68	1.44	5.34	4.54	5.35	2.46	-1.46	1.98	8.24	1.64	3.83

Увеличение времени выполнения, в процентах

Вывод №1

- Значения сильно различаются – от -11.19% до 15.38%. В целом замедление достаточно заметное – 3.83%.

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	7.19	7.64	1.63	8.55	6.07	7.49	—	-0.61	5.21	6.90	1.47	5.15
CG	4.78	4.66	1.77	4.81	3.21	3.89	6.44	-2.98	-7.02	5.52	4.14	2.66
EP	8.74	4.98	0.82	-0.15	-2.01	6.42	-11.19	-0.55	-8.75	11.31	2.14	1.07
FT	6.66	5.64	1.49	6.74	4.45	5.17	4.65	-1.24	4.84	6.76	—	4.52
IS	7.42	4.46	0.88	6.05	6.36	2.56	6.51	0.00	5.17	5.05	-0.82	3.97
LU	7.36	7.36	1.35	5.82	6.53	4.92	4.69	0.85	2.92	15.38	1.71	5.36
MG	6.50	5.90	2.14	4.75	5.78	7.39	3.63	-4.07	7.18	—	1.88	4.11
SP	5.87	4.78	1.47	6.18	5.93	4.97	—	-3.07	6.31	6.73	0.96	4.02
avg.	6.81	5.68	1.44	5.34	4.54	5.35	2.46	-1.46	1.98	8.24	1.64	3.83

Вывод №2

- Значения сильно различаются – от -11.19% до 15.38%. В целом замедление достаточно заметное – 3.83%.
- Чем больше процессов на узел – тем меньше накладные расходы.

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	7.19	7.64	1.63	8.55	6.07	7.49	—	-0.61	5.21	6.90	1.47	5.15
CG	4.78	4.66	1.77	4.81	3.21	3.89	6.44	-2.98	7.02	5.52	4.14	2.66
EP	8.74	4.98	0.82	-0.15	-2.01	6.42	-11.19	-0.55	8.75	11.31	2.14	1.07
FT	6.66	5.64	1.49	6.74	4.45	5.17	4.65	-1.24	4.84	6.76	—	4.52
IS	7.42	4.46	0.88	6.05	6.36	2.56	6.51	0.00	5.17	5.05	-0.82	3.97
LU	7.36	7.36	1.35	5.82	6.53	4.92	4.69	0.85	2.92	15.38	1.71	5.36
MG	6.50	5.90	2.14	4.75	5.78	7.39	3.63	-4.07	7.18	—	1.88	4.11
SP	5.87	4.78	1.47	6.18	5.93	4.97	—	-3.07	6.31	6.73	0.96	4.02
avg.	6.81	5.68	1.44	5.34	4.54	5.35	2.46	-1.46	1.98	8.24	1.64	3.83

Вывод №3

- Значения сильно различаются – от -11.19% до 15.38%. В целом замедление достаточно заметное – 3.83%.
- Чем больше процессов на узел – тем меньше накладные расходы.
- Влияние внешней среды существенное. И может сохраняться на длительное время
  - Причины точно не известны, но подозреваем Lustre и uptime узлов

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	7.19	7.64	1.63	8.55	6.07	7.49	—	-0.61	5.21	6.90	1.47	5.15
CG	4.78	4.66	1.77	4.81	3.21	3.89	6.44	-2.98	-7.02	5.52	4.14	2.66
EP	8.74	4.98	0.82	-0.15	-2.01	6.42	-11.19	-0.55	-8.75	11.31	2.14	1.07
FT	6.66	5.64	1.49	6.74	4.45	5.17	4.65	-1.24	4.84	6.76	—	4.52
IS	7.42	4.46	0.88	6.05	6.36	2.56	6.51	0.00	5.17	5.05	-0.82	3.97
LU	7.36	7.36	1.35	5.82	6.53	4.92	4.69	0.85	2.92	15.38	1.71	5.36
MG	6.50	5.90	2.14	4.75	5.78	7.39	3.63	-4.07	7.18	—	1.88	4.11
SP	5.87	4.78	1.47	6.18	5.93	4.97	—	-3.07	6.31	6.73	0.96	4.02
avg.	6.81	5.68	1.44	5.34	4.54	5.35	2.46	-1.46	1.98	8.24	1.64	3.83

PAPI в ручном режиме

Выводы №1 и №2

- В среднем замедление чуть больше – 4.47%.
- Когда много процессов на узел – накладные несущественны.

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	10.54	3.94	0.22	7.08	9.61	3.43	—	-3.82	18.32	7.31	0.18	5.68
CG	3.10	0.15	0.41	1.64	1.60	0.94	4.56	-8.72	5.16	2.21	0.56	1.06
EP	14.11	4.25	0.11	4.77	8.51	5.84	10.44	-3.21	13.00	14.63	2.70	6.83
FT	9.07	2.88	0.24	5.99	5.18	1.52	7.23	-2.98	8.88	5.61	—	4.36
IS	10.81	2.85	0.16	7.01	9.10	1.71	9.57	0.00	7.99	7.63	-2.02	4.98
LU	12.37	5.51	-0.25	7.33	10.83	4.61	17.44	1.12	12.60	9.93	0.07	7.42
MG	9.58	3.39	0.48	4.68	10.02	-1.49	5.47	-13.73	7.27	—	-0.60	2.51
SP	6.81	1.49	0.26	3.61	5.79	0.29	—	-6.38	9.06	5.46	0.06	2.65
avg.	9.55	3.06	0.20	5.26	7.58	2.11	9.12	-4.71	10.29	7.54	0.14	4.47

Вывод №3

- В среднем замедление чуть больше – 4.47%.
- Когда много процессов на узел – накладные несущественны.
- CG и EP – больше всего «зеленых значений». LU – наиболее подвержен влиянию шума.

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	10.54	3.94	0.22	7.08	9.61	3.43	—	-3.82	18.32	7.31	0.18	5.68
CG	3.10	0.15	0.41	1.64	1.60	0.94	4.56	-8.72	5.16	2.21	0.56	1.06
EP	14.11	4.25	0.11	4.77	8.51	5.84	10.44	-3.21	13.00	14.63	2.70	6.83
FT	9.07	2.88	0.24	5.99	5.18	1.52	7.23	-2.98	8.88	5.61	—	4.36
IS	10.81	2.85	0.16	7.01	9.10	1.71	9.57	0.00	7.99	7.63	-2.02	4.98
LU	12.37	5.51	-0.25	7.33	10.83	4.61	17.44	1.12	12.60	9.93	0.07	7.42
MG	9.58	3.39	0.48	4.68	10.02	-1.49	5.47	-13.73	7.27	—	-0.60	2.51
SP	6.81	1.49	0.26	3.61	5.79	0.29	—	-6.38	9.06	5.46	0.06	2.65
avg.	9.55	3.06	0.20	5.26	7.58	2.11	9.12	-4.71	10.29	7.54	0.14	4.47

LIKWID в ручном режиме

Выводы №1 и №2

- Наименьшее замедление в среднем – 2.78%. Наибольшее число «зеленых значений».
- Все предположения подтверждаются (число процессов на узел; поведение CG, EP, LU).

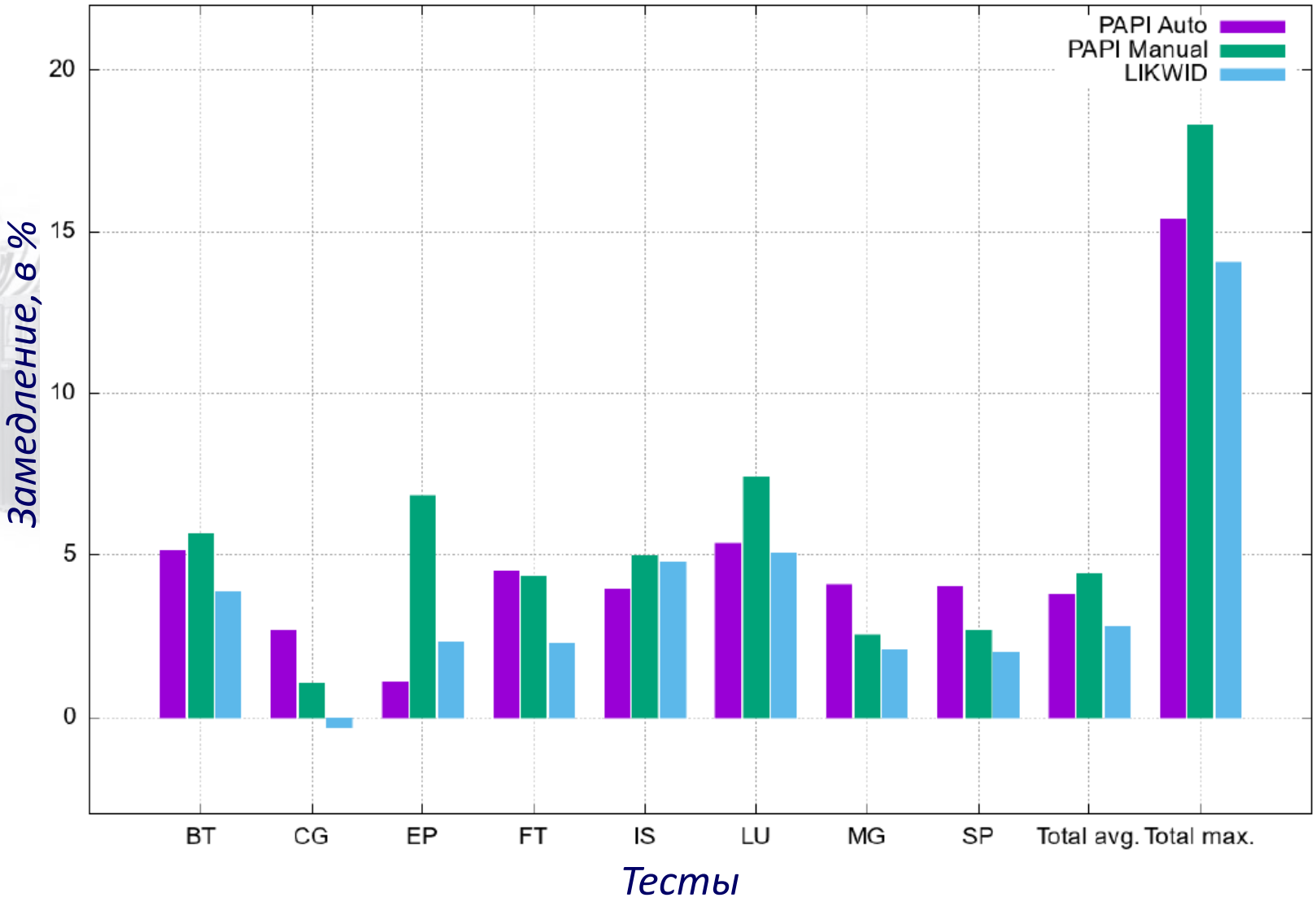
	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	10.38	3.96	0.04	5.82	8.12	2.69	—	-3.60	6.89	6.78	-2.08	3.90
CG	2.95	0.21	0.14	1.17	0.76	0.58	0.72	-6.73	-5.80	1.54	1.07	-0.31
EP	14.04	4.24	-6.49	1.00	4.33	4.47	-3.34	-2.01	-1.27	13.18	-2.68	2.32
FT	9.13	2.50	-3.66	5.58	2.04	2.07	0.52	-3.03	2.41	5.36	—	2.29
IS	10.87	3.51	1.91	5.95	9.43	0.53	9.47	0.00	7.99	7.16	-3.76	4.82
LU	12.44	4.59	-0.34	5.67	9.06	3.13	7.17	1.79	3.15	9.31	0.14	5.10
MG	9.44	3.06	-1.64	2.66	7.55	-1.16	4.99	-10.97	7.10	—	-0.31	2.07
SP	6.56	1.53	0.06	3.28	5.26	0.41	—	-6.96	4.98	5.44	-0.52	2.00
avg.	9.48	2.95	-1.25	3.89	5.82	1.59	3.25	-3.94	3.18	6.97	-1.16	2.78

Вывод №3

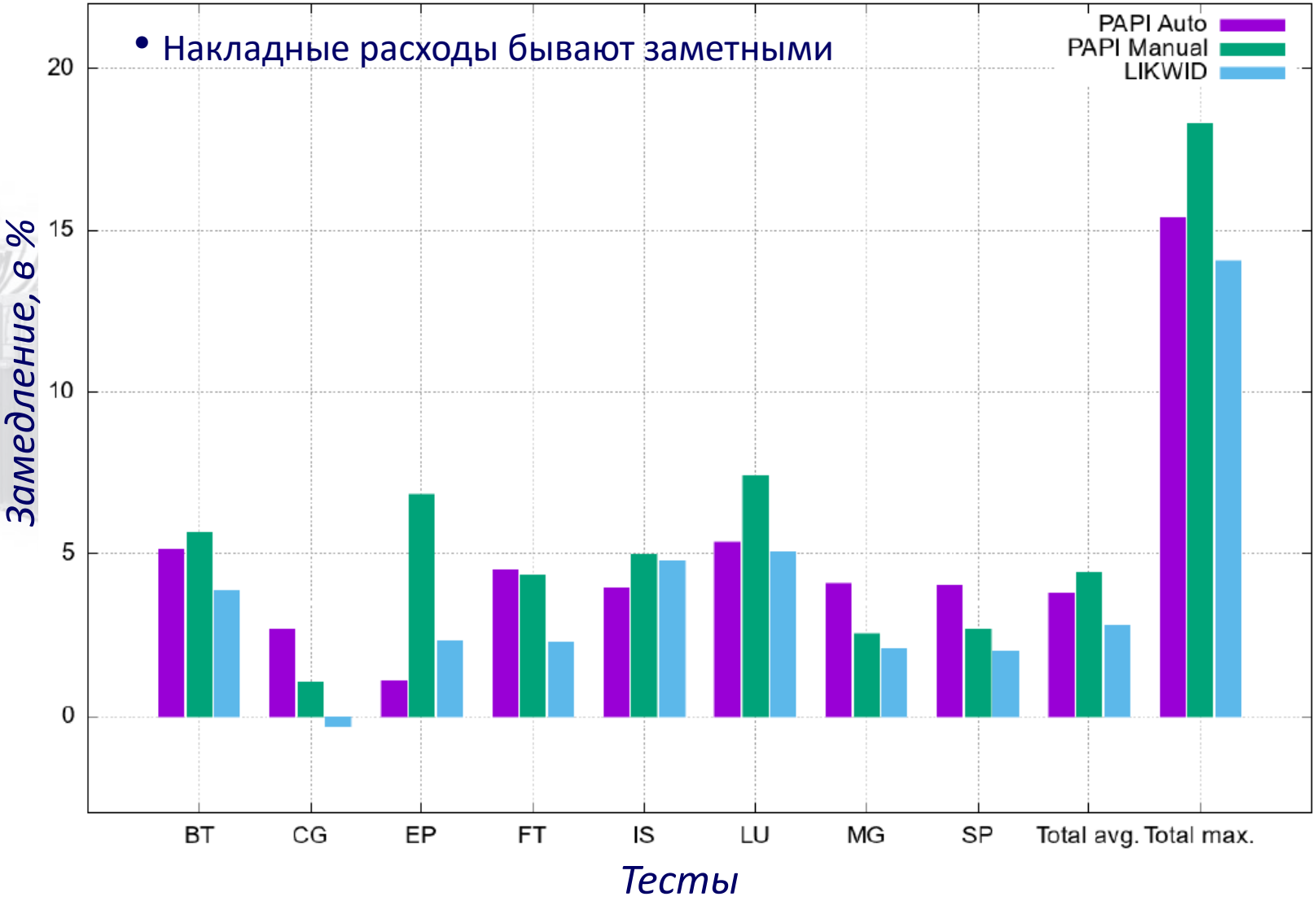
- Наименьшее замедление в среднем – 2.78%. Наибольшее число «зеленых значений».
- Все предположения подтверждаются (число процессов на узел; поведение CG, EP, LU).
- Изменение класса теста не особо влияет на накладные расходы.

	1-1-C	1-4-C	1-16-C	2-2-C	4-1-C	4-4-C	8-1-C	8-8-C	16-1-C	4-1-D	1-16-D	avg.
BT	10.38	3.96	0.04	5.82	8.12	2.69	—	-3.60	6.89	6.78	-2.08	3.90
CG	2.95	0.21	0.14	1.17	0.76	0.58	0.72	-6.73	-5.80	1.54	1.07	-0.31
EP	14.04	4.24	-6.49	1.00	4.33	4.47	-3.34	-2.01	-1.27	13.18	-2.68	2.32
FT	9.13	2.50	-3.66	5.58	2.04	2.07	0.52	-3.03	2.41	5.36	—	2.29
IS	10.87	3.51	1.91	5.95	9.43	0.53	9.47	0.00	7.99	7.16	-3.76	4.82
LU	12.44	4.59	-0.34	5.67	9.06	3.13	7.17	1.79	3.15	9.31	0.14	5.10
MG	9.44	3.06	-1.64	2.66	7.55	-1.16	4.99	-10.97	7.10	—	-0.31	2.07
SP	6.56	1.53	0.06	3.28	5.26	0.41	—	-6.96	4.98	5.44	-0.52	2.00
avg.	9.48	2.95	-1.25	3.89	5.82	1.59	3.25	-3.94	3.18	6.97	-1.16	2.78

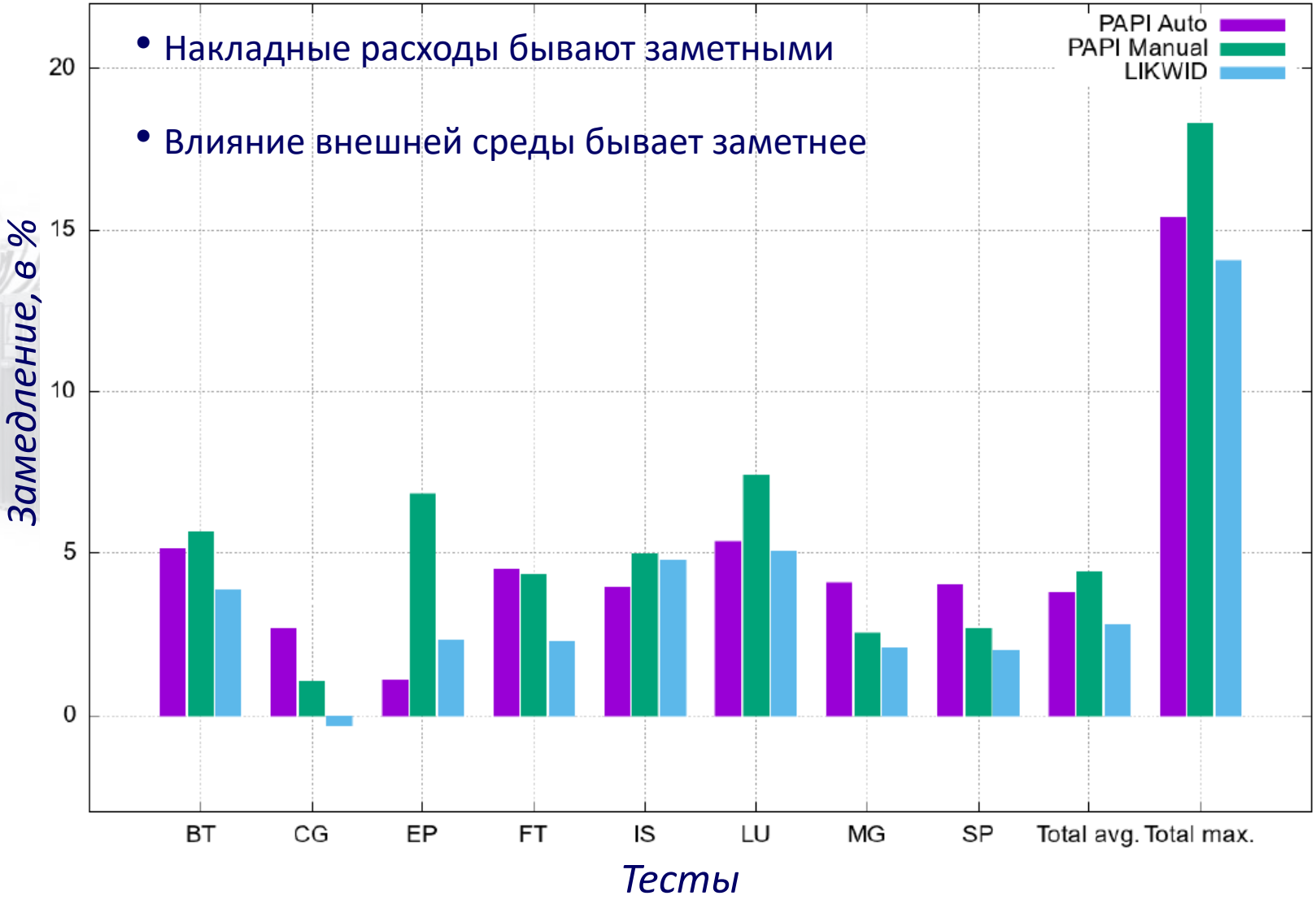
Сравнение трех вариантов



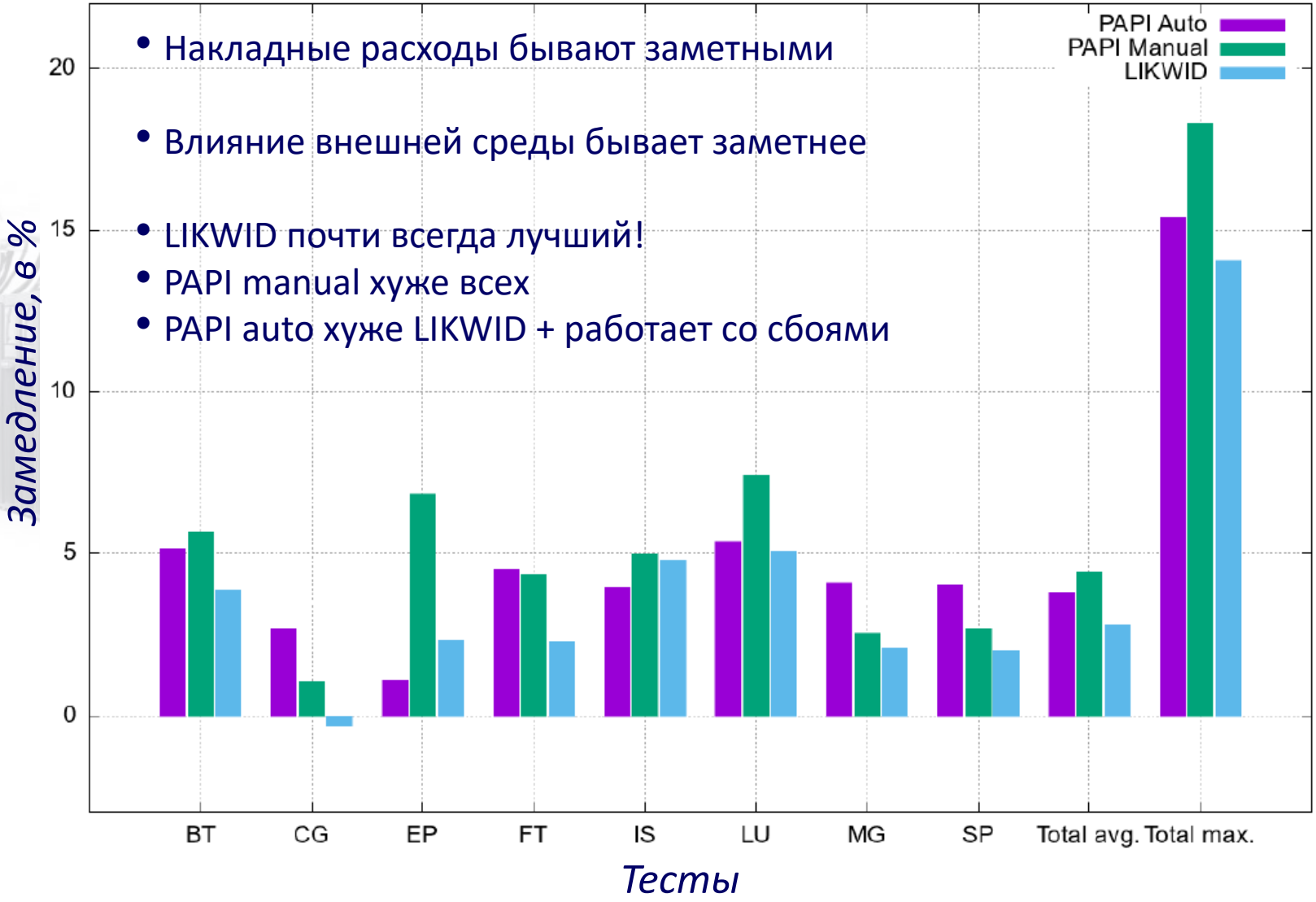
Сравнение трех вариантов



Сравнение трех вариантов



Сравнение трех вариантов



Спасибо! *Вопросы?*

Вадим Воеводин, vadim@parallel.ru

26 сентября 2022