

Программные системы: **теория и приложения**



Повышение качества видеопотока от системы технического зрения беспилотного летательного аппарата

Виталий Петрович **Фраленко**

Институт программных систем им. А. К. Айламазяна РАН, Вельково, Россия

Аннотация. В исследовании приведены результаты работы над программно-аппаратным комплексом для повышения качества видеоданных, получаемых от беспилотных летательных аппаратов. Рассмотрены задачи деконволюции отдельных кадров (удаление смазов) и стабилизации видеопотока с использованием методов машинного обучения и искусственного интеллекта. Представлены аналитические и практические результаты, позволившие подобрать решения для обработки данных от БПЛА в режиме реального времени.

Ключевые слова и фразы: БПЛА, деконволюция, стабилизация, режим реального времени, экспериментальные данные

Благодарности: Исследование выполнено за счет гранта Российского научного фонда № 21-71-10056

Для цитирования: Фраленко В.П. *Повышение качества видеопотока от системы технического зрения беспилотного летательного аппарата* // Программные системы: теория и приложения. 2023. Т. 14. № 2(57). С. 3–26.
https://psta.piras.ru/read/psta2023_2_3-26.pdf

Введение

В настоящее время особое внимание уделяется анализу и интерпретации видеопотока от беспилотных летательных аппаратов (БПЛА), которые должны обеспечивать решение задач выделения и распознавания целевых объектов, позиционирования и слежения, выполняемых с борта БПЛА, в том числе как автономно, так и с Земли, для чего жизненно необходимо иметь четкие изображения без смазов, а кадры в обрабатываемых видеопоследовательностях должны сменять друг друга без резких смещений камеры. Несмотря на удачные аэродинамические и компоновочные решения, отечественные разработки по оснащению БПЛА пока проигрывают зарубежным в части качества бортовых систем технического зрения. Для обеспечения необходимых показателей требуется создание математического, алгоритмического и программного обеспечения, способного выполнять в режиме реального времени основные функции предобработки и улучшения поступающих от БПЛА кадров.

Для проведения экспериментов с имеющимся и доработанным программным обеспечением применялся процессор общего назначения Intel Core i3 8300 и графический ускоритель вычислений NVidia GTX 1080 Ti с 11 ГБ видеопамяти. В качестве экспериментальных данных использовались снимки, полученные с мобильной камеры (набор данных GoPro¹), максимально близкие к тем, что получаются от реальных систем технического зрения БПЛА.

1. Деконволюция потока данных от БПЛА

На реальных изображениях от БПЛА присутствуют явные смазы, получившиеся из-за резкого смещения камеры во время получения того или иного кадра. При этом это смещение может быть нелинейным ввиду того, что движение БПЛА зависит от управляющих воздействий оператора, ветровой нагрузки, точности работы систем ориентирования на местности, в том числе Glonass и GPS. Все эти факторы могут накладываться на БПЛА одновременно, что приводит к тому, что положение камеры в пространстве становится нестабильным. Кроме того, объекты в процессе получения одного и того же кадра тоже могут двигаться, например, это автомобили, велосипеды, пешеходы, животные. То есть камера может некачественно фиксировать объект в кадре из-за 1) быстрого движения камеры или самого объекта; 2) неоптимальных настроек камеры (например,

¹Набор данных GoPro для удаления размытия (*Papers with Code*^{url}) состоит из 3214 размытых изображений размером 1280 × 720, которые разделены на 2103 обучающих изображения и 1111 тестовых изображений. Набор данных состоит из пар реалистичного размытого изображения и соответствующего наземного изображения, полученного с помощью высокоскоростной камеры.

поставлена слишком большая выдержка, матрица очень медленная или перегрета). Деконволюция (от англ. «deconvolution») – способ компенсации таких факторов, приводящих к тому, что кадры видеопотока становятся смазанными. Предлагается использовать для деконволюции методы математической статистики и искусственные нейронные сети (ИНС), обучаемые с помощью общедоступных наборов данных.

Примеры обрабатываемых изображений приведены на рисунках 1– 2.

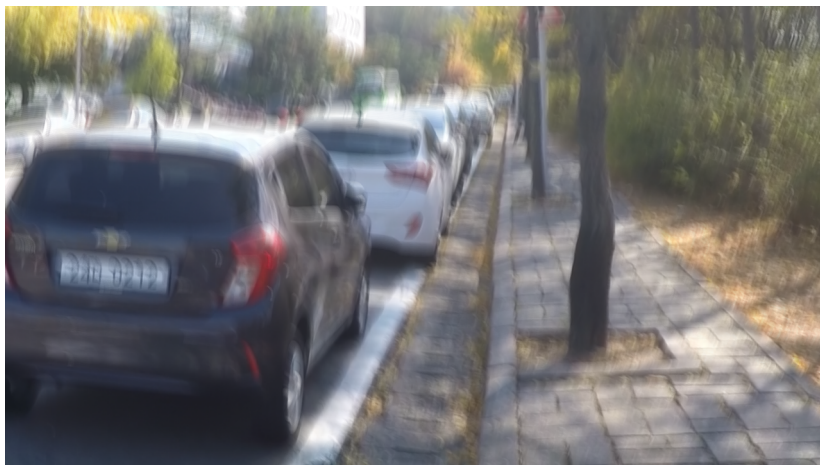


Рисунок 1. Пример 1 исходного изображения

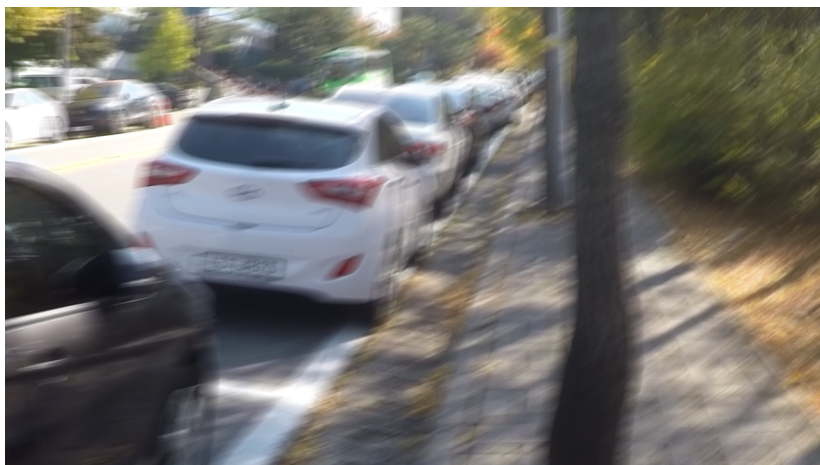
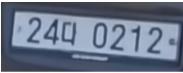
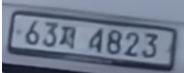
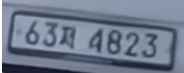
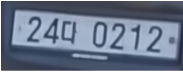
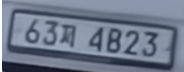
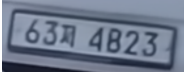
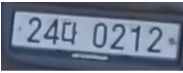
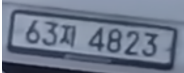
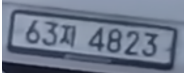
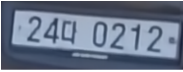
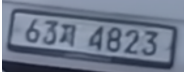
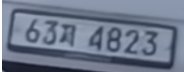
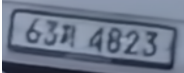
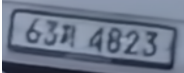
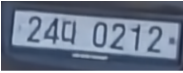
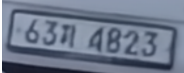
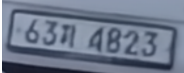
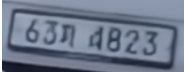
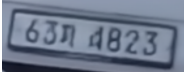
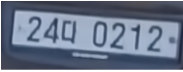
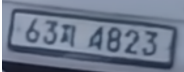
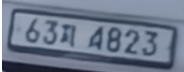


Рисунок 2. Пример 2 исходного изображения

В таблице 1 перечислены выявленные в результате аналитического исследования алгоритмические и программные решения, предназначенные для автоматизированного устранения смазов. Приводятся сведения по сред-

ТАБЛИЦА 1. Результаты деконволюции

Метод	Время, с.	Фрагмент 1		Фрагмент 2	
		PSNR	SSIM	PSNR	SSIM
<i>Test-time Local Converter (TLC)</i> ^{URL} [1]	13.5	 33.012	 0.939	 31.186	 0.929
<i>Multi-Axis MLP</i> ^{URL} [2]	19.5	 30.556	 0.917	 31.748	 0.934
<i>EFNet</i> ^{URL} [3]	1.10	 35.184	 0.948	 35.695	 0.962
<i>Learning degradation</i> ^{URL} [4]	3.60	 32.549	 0.932	 31.739	 0.934
<i>Deep Generalized Unfolding (DGU)</i> ^{URL} [5]	5.80	 22.582	 0.735	 29.923	 0.903
<i>NAFNet width32</i> ^{URL} [6]	5.00	 31.789	 0.918	 30.000	 0.916
<i>Stripformer</i> ^{URL} [7]	0.03	 28.060	 0.876	 30.222	 0.916
<i>Uformer</i> ^{URL} [8]	9.40	 32.296	 0.928	 31.477	 0.929

нему времени обработки отдельных кадров с помощью соответствующего

решения и результаты обработки в виде фрагментов целевых областей. Отметим, что все актуальные решения реализованы в виде программного кода, использующего для вычислений графические процессоры.

В рамках настоящей работы решалась подзадача восстановления областей видеокadra с присутствующим номером автомобиля. БПЛА активно используются в мире для автоматизированного обследования и фиксации автомобильных номеров, поэтому подобная подзадача является актуальной. В данном случае используются изображения с автомобильными номерами Республики Корея, третий символ в номере автомобиля это буквенный префикс, а не цифра, как может показаться. Для численной оценки были получены соответствующие отношения сигнала к шуму (Peak Signal-to-Noise Ratio, PSNR) и значения величин структурных подобий (Structure SIMilarity, SSIM), использовалась программная библиотека [skimage.metrics](https://github.com/ajkjavaher/skimage-metrics)^[URL]. Выполнено сравнение полученных изображений с изображениями без смазов (подгруппа снимков «sharp» из набора данных GoPro). Для поврежденных смазами оригинальных изображений автомобильных номеров (см. рисунки 1– 2) PSNR и SSIM соответственно равны следующим величинам: 18.42 и 0.57 (рисунок 1); 18.79 и 0.56 (рисунок 2). При полном совпадении изображений значение $SSIM = 1$, для абсолютно разных изображений $SSIM = 0$.

Экспериментальные исследования, проведенные с использованием этих и других изображений из набора данных GoPro, показали, что нейронная сеть Stripformer выполняет обработку в режиме реального времени, при этом получаются достаточно качественные выходные изображения. Stripformer имеет архитектуру, основанную на идее трансформеров. «Визуальный трансформер» состоит из токенизатора, трансформера и проектора, более подробное описание принципов его работы можно найти в исследовании [9]. Stripformer строит внутренние и внешние ленточные токены (strip tokens) в процессе перерасчета информативных признаков изображения в горизонтальных и вертикальных направлениях, что позволяет выявлять зашумленные фрагменты с различной ориентацией (направлением распространения) смаза. Нейронная сеть совмещает внешние и внутренние ленточные слои внимания (attention layers), чтобы определить степень и характер размытия. В дополнение к обнаружению

специфических для области размытых паттернов различных ориентаций и размеров, Stripformer успешно работает даже без данных о динамике сцены и не требует огромного массива данных для обучения. Следует отметить высокое качество работы нейронной сети EFNet.

Пример работы Stripformer показан на рисунке 3. Обученная оригинальная модель требует для своей работы наличия видеокарты с не менее чем 8 ГБ видеопамати. В рамках настоящего исследования выполнена небольшая модификация программного кода Stripformer, связанная с особенностями запуска модели на заявленном процессоре общего назначения с использованием библиотеки *PyTorch*^{URL}, однако она работает неприемлемо долго, порядка 90 секунд на каждый кадр, что в 3000 раз дольше, чем на используемой видеокarte.



Рисунок 3. Результат удаления смазов на снимке с помощью Stripformer

Таким образом, предложенное решение позволяет в режиме реального времени получать данные от БПЛА, очищенные от смазов.

2. Стабилизация видеопотока

При стабилизации видеопотока кадр обрезается со всех сторон, а оставшиеся части кадра используются для компенсации движения

и формирования итогового изображения. Эта компенсация обычно выполняется с помощью ИНС и статистических методов на основе предыдущих кадров, для чего поддерживается буфер определенного размера.

В качестве тестовых в настоящей работе использовались выборки по 100 кадров, восстановленные с помощью Stripformer и подаваемые со скоростью 24 кадра в секунду.

Далее перечислены выявленные в результате исследования алгоритмические и программные решения, предназначенные для автоматизированной стабилизации видеопотока от БПЛА. Приводится среднее время обработки всего видеофайла. Каждое решение снабжено комментарием, отражающим приобретенный опыт его использования, некоторые реализованы с поддержкой графических ускорителей вычислений.

- **VidGear**^{URL} [10]. 3.91 секунды. Рамки кадров обрезаются с помощью отступа от границ и зума. Требуется инициализация стабилизатора, для чего используется объем входных данных порядка 24 кадров. Время подготовки стабилизатора к работе составляет порядка 0.70 с., то есть оно выполняется в фоновом режиме по мере поступления кадров на 30% быстрее того, как кадры поступают на обработку.
- **FuSta: Hybrid Neural Fusion**^{URL} [11]. 1200 секунд. Используется графический ускоритель. На первом шаге запускается скрипт с генерацией данных оптического потока (*python main.py* с параметрами), который попутно выполняет стабилизацию видео, но без отрезания краев, зума и пр. Полная обработка 100 кадров заняла 5 минут. Итоговый стабилизированный файл получился с разрешением 960x576 пикселей. На втором шаге запускается основная программа, использующая собранные данные оптического потока, *python run_FuSta.py* с опцией *-temporal_width 18* вместо исходной *-temporal_width 41*. Иначе графическому ускорителю не хватает имеющихся 11 ГБ видеопамати. Оригинальная библиотека содержала уже неподдерживаемые коды нейронной сети flownet2, поэтому пришлось переписать их. Результат – набор изображений в исходном разрешении. Обработка выполняется за 15 минут. Видео, получаемое из этих итоговых кадров плавное, однако, в отличие от результатов, полученных с помощью скрипта *python main.py*, на границах объектов «автомобиль» и «дерево» имеются артефакты (см. рисунок 4).



Рисунок 4. Артефакты, которые получены при работе с FuSta

- *MeshFlow: Minimum Latency Online Video Stabilization*^{URL} [12]. 1260 секунд (см. рисунок 5). Проведены эксперименты с двумя профилями



Рисунок 5. Оригинальное изображение до стабилизации и артефакт с наклоном автомобиля при использовании профиля Constant High метода MeshFlow

настроек. Профиль Original: отличное качество без рамок, кадр

обрезается автоматически. Профиль Constant High: обнаружены артефакты в виде неестественно наклоненных объектов, например, автомобилей.

- *Video Stabilization with L1 optimal camera paths*^{URL} [13]. 0.50 секунды. Результирующее видео формируется как будто волнами. Камера то наезжает на объекты в кадре, то возвращается обратно, что выглядит неестественным.
- *Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths*^{URL} [13] (используется дополнительная предобработка изображений). 120 секунд. Итоговое изображение существенно обрезается при применении достаточно высоких уровней стабилизации. Например, при пороге в 40 пикселей изображение 1280x720 уменьшается до 1180x620. Использование меньшего порога дает недостаточный уровень стабилизации видео. Обработка разбита на несколько этапов: 2 минуты на предобработку, затем быстрая стабилизация с использованием полученных данных.
- *Video stabilization using homography transform*^{URL} [14]. 4.18 секунды. Итоговое изображение существенно обрезается, особенно при высоких коэффициентах стабилизации. Наилучшие результаты получены на realtime-стабилизации, при значении параметра *amount=20*.
- *DUT: Learning Video Stabilization*^{URL} [15]. 32 секунды. Используется графический ускоритель. Метод опирается на несколько предобученных нейронных сетей. В основе – самообучение на нестабильных видеоданных. Изображение автомобиля на итоговом видео немного отклоняется влево-вправо (не так существенно, как на рисунке 5), однако в целом получившееся видео можно назвать качественным.
- *Deep Iterative Frame Interpolation for Full-frame Video Stabilization*^{URL} [16]. 100 секунд. Используется графический ускоритель. Обработка выполняется итеративно, число итераций регулируется параметром *n_iter*, по умолчанию *n_iter* = 3. Итоговый видеопоток имеет неустойчивый характер, камера «прыгает» от кадра к кадру, общее качество создаваемого видеопотока низкое.
- *PWStableNet*^{URL} [17]. 8 секунд. Используется графический ускоритель. Результирующее видео с подергиваниями.

Наилучшие результаты в задаче стабилизации получены с использованием библиотеки VidGear, в которой накапливается массив особых точек на заданном числе кадров с дальнейшим его обновлением в режиме реального времени. Накопленные данные используются для оценки

движения камеры БПЛА с дальнейшей компенсацией ее неустойчивости. Реализация соответствует подходу, используемому в библиотеке *OpenCV*^{[URL](#)}. При этом в целом обработка происходит в конвейерном режиме.

Заключение

Проведенные исследования позволили получить решение двух важнейших задач улучшения видеопотока от БПЛА: удаление смазов и стабилизация видео. Нейронная сеть Stripformer кадрами удаляет смазы, затем обработанные данные сразу попадают в буфер модуля стабилизации на базе библиотеки VidGear, с помощью которого осуществляется стабилизация видеопотока. Итоговое решение позволяет осуществлять полный цикл обработки видеопотока с задержкой не более чем на 0.03 секунды за счет применения конвейерного метода обработки данных. Обработка выполняется в режиме реального времени.

Список литературы

- [1] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, Xin Lu *Improving image restoration by revisiting global information aggregation*, ECCV 2022: Computer Vision – ECCV 2022 (October 23–27, 2022, Tel Aviv, Israel), Lecture Notes in Computer Science.– vol. **13678**, Cham: Springer.– 2022.– Pp. 330–351 .  [arXiv:2112.04491](#) ↑6
- [2] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, Yinxiao Li *MAXIM: multi-axis MLP for image processing // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.– 2022.– 34 pp.  [arXiv:2201.02973](#) ↑6
- [3] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, Luc Van Gool *Event-based fusion for motion deblurring with cross-modal attention*, ECCV 2022: Computer Vision – ECCV 2022 (October 23–27, 2022, Tel Aviv, Israel), Lecture Notes in Computer Science.– vol. **13678**, Cham: Springer.– Pp. 412–428; 2023.– 17 pp.  [arXiv:2112.00167](#) ↑6
- [4] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, Hongsheng Li *Learning degradation representations for image deblurring // Computer Vision – ECCV 2022*, ECCV 2022, Lecture Notes in Computer Science, eds. Avidan S., Brostow G. Cissé M. Farinella G.M. Hassner T.; 2022.– 18 pp.  [arXiv:2208.05244](#) ↑6

- [5] Chong Mou, Qian Wang, Jian Zhang *Deep generalized networks for image restoration // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*– 2022.– Pp. 17399–17410.– 12 pp. [arXiv:2204.13348](#) [doi](#) [↑6](#)
- [6] Chen L., Chu X., Zhang X., Sun J. *Simple baselines for image restoration // Computer Vision – ECCV 2022, ECCV 2022, Lecture Notes in Computer Science*, eds. Avidan S., Brostow G., Cissé M., Farinella G.M., Hassner T..– 2022; 21 pp. [arXiv:2204.04676](#) [doi](#) [↑6](#)
- [7] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, Chia-Wen Lin *Stripformer: strip transformer for fast image deblurring. // Computer Vision – ECCV 2022, ECCV 2022, Lecture Notes in Computer Science.*– vol. **13679**, eds. Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T..– 2022; 17 pp. [arXiv:2204.04627](#) [doi](#) [↑6](#)
- [8] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, Houqiang Li *Uformer: a general U-shaped transformer for image restoration*, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).– 2022.– 17 pp. [arXiv:2106.03106](#) [doi](#) [↑6](#)
- [9] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, Peter Vajda *Visual transformers: token-based image representation and processing for computer vision.*– 2020.– 12 pp. [arXiv:2006.03677](#) [doi](#) [↑7](#)
- [10] Thakur A., Papakipos Z., Clauss C., Hollinger C., Andolina I. M., Boivin V., enarche-ahn, freol35241, Lowe B., Schoentgen M., Bouckenoghe R. *abhiTronix/vidgear: VidGear v0.2.6.*– 2022. [doi](#) [URL](#) [↑9](#)
- [11] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, Jia-Bin Huang *Hybrid neural fusion for full-frame video stabilization*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).– 2021.– Pp. 2299–2308. [arXiv:2102.06205](#) [doi](#) [↑9](#)
- [12] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, Bing Zeng *MeshFlow: minimum latency online video stabilization*, ECCV 2016: Computer Vision – ECCV 2016 (October 11-14, 2016, Amsterdam, The Netherlands), Lecture Notes in Computer Science.– vol. **9910**.– 2016.– Pp. 800–815. [doi](#) [URL](#) [↑10](#)
- [13] Grundmann M., Kwatra V., Essa I. *Auto-directed video stabilization with robust L1 optimal camera paths // CVPR '11: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (20–25 June 2011, Colorado Springs, CO, USA).*– IEEE Computer Society.– 2011.– ISBN 978-1-4577-0394-2.– Pp. 225–232. [doi](#) [↑11](#)

- [14] Grundmann M., Kwatra V., Essa I. *Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths* // *CVPR '11: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (20–25 June 2011, Colorado Springs, CO, USA).– IEEE Computer Society.– 2011.– Pp. 225–232. doi ↑11
- [15] Yufei Xu, Jing Zhang, Stephen J. Maybank, Dacheng Tao *DUT: learning video stabilization by simply watching unstable videos* // *IEEE Transactions on Image Processing*.– 2022.– Vol. **31**.– Pp. 4306–4320 . doi arXiv 2011.14574 ↑11
- [16] Jinsoo Choi, In So Kweon *Deep iterative frame interpolation for full-frame video stabilization* // *ACM Transactions on Graphics*.– 2019.– Vol. **39**.– No. 1.– id. 4.– 9 pp. doi arXiv 1909.02641 ↑11
- [17] Wang M., Yang G. -Y., Lin J. -K., Zhang S. -H., Shamir A., Lu S. -P., Hu S. -M. *Deep online video stabilization with multi-grid warping transformation learning* // *IEEE Transactions on Image Processing*.– 2019.– Vol. **28**.– No. 5.– Pp. 2283–2292 . doi arXiv 1909.02641 ↑11

Поступила в редакцию 11.03.2023;
одобрена после рецензирования 23.03.2023;
принята к публикации 28.03.2023;
опубликована онлайн 07.07.2023.

Рекомендовал к публикации

д.ф.-м.н. А. М. Елизаров

Информация об авторе:



Виталий Петрович Фраленко

Кандидат технических наук, ведущий научный сотрудник ИЦМС ИПС им. А.К. Айламазяна РАН. Область научных интересов: интеллектуальный анализ данных и распознавание образов, искусственный интеллект и принятие решений, параллельно-конвейерные вычисления, сетевая безопасность, диагностика сложных технических систем, графические интерфейсы.



0000-0003-0123-3773

e-mail: alarmod@pereslavl.ru

Автор заявляет об отсутствии конфликта интересов.



Improving quality of video stream from the unmanned aerial vehicle technical vision system

Vitaly Petrovich **Fralenko**

Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

Abstract. The study contains the results of work on the software and hardware complex to improve the quality of video data obtained from unmanned aerial vehicles. Including the tasks of independent video-flow images deconvolution (motion blur removal) and stabilization of the video stream using machine learning and artificial intelligence methods. Analytical and practical results are presented that allowed to choose solutions for processing data from UAVs in real time.

Key words and phrases: UAV, deconvolution, stabilization, real time, experimental data

2020 *Mathematics Subject Classification:* 68T07; 68T40, 68U10

Acknowledgments: This work was financially supported by the Russian Science Foundation, project 21-71-10056

For citation: Vitaly P. Fralenko. *Improving quality of video stream from the unmanned aerial vehicle technical vision system.* Program Systems: Theory and Applications, 2023, 14:2(57), pp. 3–26. https://psta.psiras.ru/read/psta2023_2_3-26.pdf

Introduction

Currently, special attention is paid to the analysis and interpretation of video stream from unmanned aerial vehicles (UAVs), which should ensure the solution of the problems of the allocation and recognition of target objects, positioning and tracking, performed from the UAV's board, including autonomously mode and with using hardware and software on Earth, which is vital to have clear images without “motion blur” effect, and frames in the processed video stream should replace each other without sharp shifts of the camera. Despite successful aerodynamic and layout solutions, domestic developments for equipping UAVs are still losing foreign in terms of the on-board technical vision systems quality. To ensure the necessary indicators, the creation of mathematical, algorithmic and software is required, capable of performing real-time mode of the main functions of exporting and improving the frames from UAVs.

To conduct experiments with existing and modified software, was used a general purpose processor Intel Core i3 8300 and a graphic computing accelerator NVIDIA GTX 1080 Ti with 11 GB of video memory. As experimental data, images of 1280x720 pixels received from a mobile chamber (Dataset Gopro¹), as close as possible from real technical vision systems of the UAV.

1. Deconvolution of UAV data stream

Real images from UAVs contain apparent motion blur caused by the sharp camera displacement due to operator control inputs, wind loads, and the operation of orientation systems on the ground, including GLONASS and GPS. All these factors can occur simultaneously, which destabilizes the camera's position in space. In addition, objects such as cars, bicycles, pedestrians, and animals can also move during the capture of the same frame. Therefore, the camera may inadequately capture an object in the frame due to 1) rapid camera or object movement, or 2) suboptimal camera settings (for example, an excessively long exposure time, or a very slow or

¹The GoPro dataset for deblurring (*Papers with Code*[®]) consists of 3,214 blurred images with the size of 1,280×720 that are divided into 2,103 training images and 1,111 test images. The dataset consists of pairs of a realistic blurry image and the corresponding ground truth shapr image that are obtained by a high-speed camera.

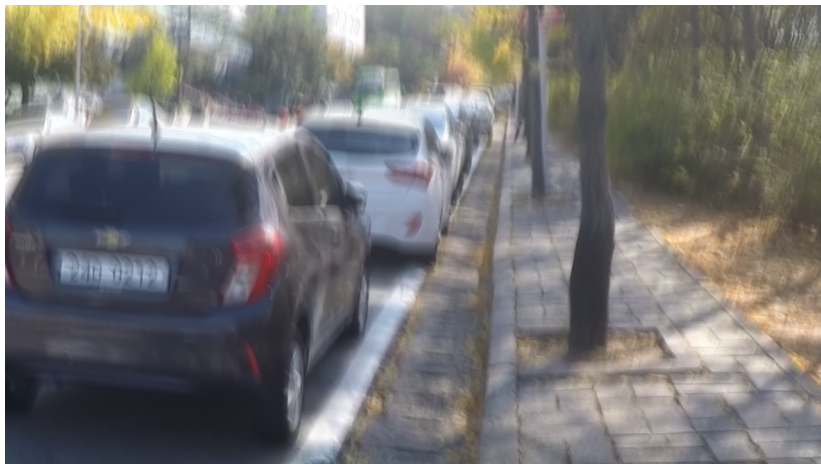


FIGURE 1. Original image example 1

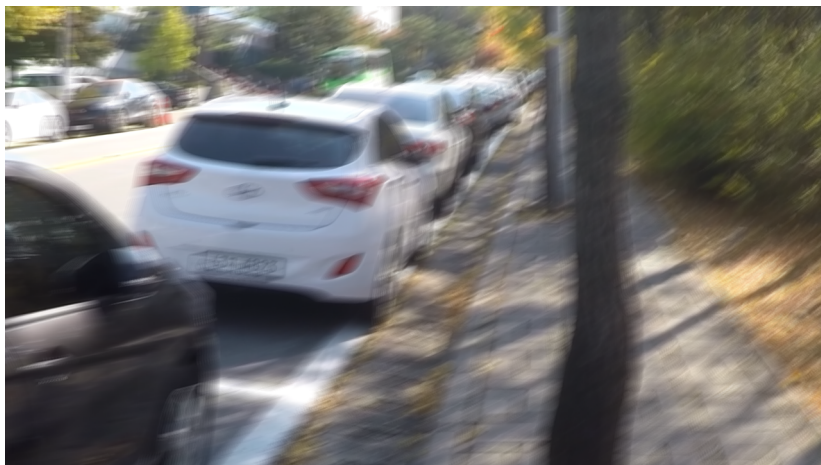


FIGURE 2. Original image example 2

overheated matrix). Deconvolution is a method of compensating for these factors that cause video stream frames to become blurred. We propose to use for deconvolution the methods of mathematical statistics and artificial neural networks (ANNs) trained with publicly available datasets.

Examples of processed images are shown in Figure 1–2.

TABLE 1. Deconvolution results

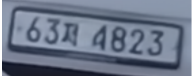
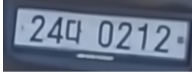
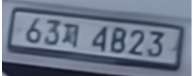
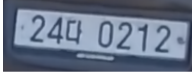
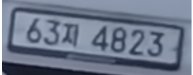
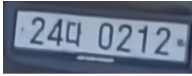
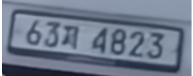
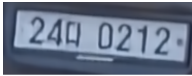
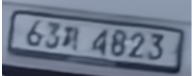
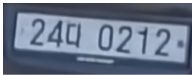
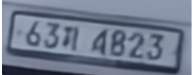
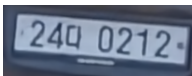
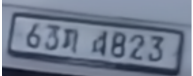
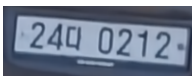
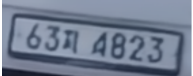
Method	Processing time, s.	Fragment 1		Fragment 2	
		PSNR	SSIM	PSNR	SSIM
<i>Test-time Local Converter (TLC)</i> ^{URL} [1]	13.5	 33.012	 0.939	31.186	0.929
<i>Multi-Axis MLP</i> ^{URL} [2]	19.5	 30.556	 0.917	31.748	0.934
<i>EFNet</i> ^{URL} [3]	1.10	 35.184	 0.948	35.695	0.962
<i>Learning degradation</i> ^{URL} [4]	3.60	 32.549	 0.932	31.739	0.934
<i>Deep Generalized Unfolding (DGU)</i> ^{URL} [5]	5.80	 22.582	 0.735	29.923	0.903
<i>NAFNet width32</i> ^{URL} [6]	5.00	 31.789	 0.918	30.000	0.916
<i>Stripformer</i> ^{URL} [7]	0.03	 28.060	 0.876	30.222	0.916
<i>Uformer</i> ^{URL} [8]	9.40	 32.296	 0.928	31.477	0.929

Table 1 lists algorithmic and software solutions designed for the automated motion blur removal. The information is provided as individual

frames average processing time (with using the appropriate solution) and original processing results in the form of target areas fragments. It should be noted that all current solutions are implemented as software code using graphics processors for calculations.

This study addressed to the subtask of restoring areas of a video frame with a car number visible. UAVs are actively used worldwide for automated inspection and registration of car numbers, making this subtask relevant. In this case, images with car numbers from the Republic of Korea are used, the third symbol in the car room is a letter prefix, not a figure, as it might seem. For a numerical assessment, the corresponding relationships of the Peak Signal-to-Noise Ratio (PSNR) and the values of the Structure Similarity (SSIM) were obtained, the *skimage.metrics*^{URL} program library was used. A comparison of the obtained images with images without motion blur (subgroup images “sharp” from the GoPro dataset) was performed. The metrics for comparing PSNR and SSIM image pairs for the damaged original blurred images of car numbers (see Figure 1–2), respectively, are as follows: 18.42 and 0.57 (Figure 1); 18.79 and 0.56 (Figure 2). With a complete coincidence of images, the value of $SSIM = 1$, for completely different images $SSIM = 0$.

Experimental studies conducted using these and other images from the GoPro dataset have shown that the Stripformer neural network performs real-time processing while producing high-quality output images. Stripformer has an architecture based on the transformers idea. The “visual transformer” consists of a tokenizer, transformer, and projector, and a more detailed description of its operating principles can be found in the study [9]. Stripformer constructs internal and external strip tokens during the recalculation of informative image features in horizontal and vertical directions, allowing it to identify noisy fragments with different orientations (direction of blur propagation). The neural network combines external and internal strip attention layers to determine the degree and character of a blur. In addition to detecting specific blurred patterns of different orientations and sizes, Stripformer works successfully even



FIGURE 3. Result of removing motion blur in the image using Stripformer

without data on the dynamics of the scene and does not require a huge amount of training data. It is worth noting the high quality of the EFNet neural network.

An example of Stripformer's model result is shown in Figure 3. A trained original model requires a video card with at least 8 GB of video memory for its work. Within the framework of this study, a small modification of the StripFormer software code has been performed, related to the features of the model launch on the stated general purpose processor using the *PyTorch*^{URL} library, but it works unacceptably for a long time, about 90 seconds for each frame, which is 3000 times longer than on the used video card.

Thus, the proposed solution allows for obtaining data from UAVs without motion blur in real-time.

2. Video stream stabilization

During video stream stabilization, the frame is cropped on all sides, and the remaining parts of the frame are used to compensate for motion and

generate the final image. This compensation is usually performed using ANN and statistical methods based on previous frames, which requires maintaining a buffer of a certain size.

In this study, test sets of 100 frames were used, restored using Stripformer and fed at a rate of 24 frames per second.

Next are listed the algorithmic and software solutions intended for automated UAV video stream stabilization. The average processing time for the entire video file is provided. Each solution is accompanied by a comment reflecting the acquired experience of its use, some are implemented with support of a Graphics Processing Units.

- *VidGear*^{URL} [10]. 3.91 seconds. The frame borders are cropped using a border offset and scaling. Initialization of the stabilizer is required, with input data of about 24 frames used. The time required to prepare the stabilizer for operation is approximately 0.70 s., that is, it is performed in the background as the frames turn 30% faster than the frames enter to the processing.
- *FuSta: Hybrid Neural Fusion*^{URL} [11]. 1200 seconds. A graphics accelerator is used. At the first step, a script is launched with the generation of optical stream data (*python main.py* with parameters), which simultaneously performs video stabilization, but without cutting the edges, zoom, etc. Complete processing of 100 frames took 5 minutes. The final stabilized file turned out with a resolution of 960x576 pixels. At the second step, the main program is launched using the collected data of the optical flow, *python run_FuSta.py* with the *-temporal_width 18* instead of the original *-temporal_width 41*. Otherwise, the graphic accelerator is not enough of the available 11 GB video memory. The original library contained the already unsupported codes of the Flownet2 neural network, so its software codes were rewritten. Result — a set of images in the initial resolution. Processing is performed in 15 minutes. The video obtained from these final frames is smooth, however, in contrast to the results obtained using the *python main.py* script, there are artifacts on the boundaries of objects “car” and “tree” (see Figure 4).



FIGURE 4. Artifacts that received when working with Fusta



FIGURE 5. Original image to stabilization and an artifact with a car inclination using the Constant High profile of the Meshflow method

- *MeshFlow: Minimum Latency Online Video Stabilization*^{URL} [12]. 1260 seconds (see Figure 5). Experiments were conducted with two settings

profiles. The Original profile provides excellent quality without borders, and the frame is automatically cropped. The Constant High profile has artifacts in the form of unnaturally tilted objects, such as cars.

- *Video Stabilization with L1 optimal camera paths*^{URL} [13]. 0.50 seconds. The resulting video is “wavy”. The camera sometimes moves towards objects in the frame and then back, what looks unnatural.
- *Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths*^{URL} [13] (additional image preprocessing is used). 120 seconds. The resulting image is significantly cropped when high levels of stabilization are used. For example, at a threshold of 40 pixels, an image of 1280x720 is cropped at the edges to 1180x620. Using a lower threshold gives an insufficient level of video stabilization. Processing is divided into two stages: 2 minutes for preprocessing, followed by fast stabilization using the obtained data.
- *Video stabilization using homography transform*^{URL} [14]. 4.18 seconds. The resulting image is significantly cropped, especially at high stabilization coefficients. The best results are obtained with realtime-stabilization, with the *amount* parameter set to 20.
- *DUT: Learning Video Stabilization*^{URL} [15]. 32 seconds. A graphics accelerator is used. The method relies on several pre-trained neural networks. The approach is based on self-learning on unstable video data. The car image in the resulting video is slightly offset left and right (not as essential as in the. 5), but the overall resulting video can be considered of good quality.
- *Deep Iterative Frame Interpolation for Full-frame Video Stabilization*^{URL} [16]. 100 seconds. Processing is performed iteratively, with the number of iterations controlled by the *n_iter* parameter, which is set to 3 by default. The resulting video stream is unstable, and the overall quality of the resulting video stream is low.
- *PWStableNet*^{URL} [17]. 8 seconds. A graphics accelerator is used. The resulting video has jerky movements.

The best stabilization results were achieved using the VidGear library, which accumulates an array of key points over a specified number of frames

with subsequent real-time updating. The accumulated data is used to estimate the motion of the UAV camera with compensation for its jerks. The implementation corresponds to the approach used in the *OpenCV*^{[URL](#)} library. The overall processing occurs in a pipeline mode, as deconvolution and stabilization are performed in parallel.

Conclusion

The conducted research allowed to obtain a solution to two of the most important tasks of improving the video stream from UAVs: removing blurs and stabilizing the video. The Stripformer neural network removes motion blur frame by frame, and then the processed data immediately enters the buffer of the stabilization module based on the VidGear library, which stabilizes the video stream. The final solution allows for the full cycle of video stream processing with a delay of no more than 0.03 seconds due to the use of a pipeline method for data processing. Processing is carried out in real-time mode.

References

- [1] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, Xin Lu. “Improving image restoration by revisiting global information aggregation”, ECCV 2022: Computer Vision – ECCV 2022 (October 23–27, 2022, Tel Aviv, Israel), Lecture Notes in Computer Science, vol. **13678**, Springer, Cham, 2022, pp. 330–351 .  [arXiv](#) [2112.04491](#) ↑¹⁸
- [2] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, Yinxiao Li. “MAXIM: multi-axis MLP for image processing”, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 34 pp.  [arXiv](#) [2201.02973](#) ↑¹⁸
- [3] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, Luc Van Gool. “Event-based fusion for motion deblurring with cross-modal attention”, ECCV 2022: Computer Vision – ECCV 2022 (October 23–27, 2022, Tel Aviv, Israel), Lecture Notes in Computer Science, vol. **13678**, Springer, Cham, pp. 412–428; 2023, 17 pp.  [arXiv](#) [2112.00167](#) ↑¹⁸
- [4] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, Hongsheng Li. “Learning degradation representations for image deblurring”, *Computer Vision – ECCV 2022*, ECCV 2022, Lecture Notes in Computer Science, eds. Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T.; 2022, 18 pp.  [arXiv](#) [2208.05244](#) ↑¹⁸

- [5] Chong Mou, Qian Wang, Jian Zhang. “Deep generalized unfolding networks for image restoration”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17399–17410, 12 pp. [arXiv](#)  [2204.13348](#)  [↑18](#)
- [6] L. Chen, X. Chu, X. Zhang, Sun J.. “Simple baselines for image restoration”, *Computer Vision – ECCV 2022*, ECCV 2022, Lecture Notes in Computer Science, eds. Avidan S., Brostow G., Cissé M., Farinella G.M., Hassner T., 2022; 21 pp.  [arXiv](#)  [2204.04676](#) [↑18](#)
- [7] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, Chia-Wen Lin. “Stripformer: strip transformer for fast image deblurring.”, *Computer Vision – ECCV 2022*, ECCV 2022, Lecture Notes in Computer Science, vol. **13679**, eds. Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., 2022; 17 pp.  [arXiv](#)  [2204.04627](#) [↑18](#)
- [8] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, Houqiang Li. “Uformer: a general U-shaped transformer for image restoration”, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 17 pp.  [arXiv](#)  [2106.03106](#) [↑18](#)
- [9] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, Peter Vajda. *Visual transformers: token-based image representation and processing for computer vision*, 2020, 12 pp. [arXiv](#)  [2006.03677](#)  [↑19](#)
- [10] A. Thakur, Z. Papakipos, C. Clauss, C. Hollinger, I. M. Andolina, V. Boivin, enarche-ahn, freol35241, B. Lowe, M. Schoentgen, R. Bouckennooghe. “abhiTronix/vidgear: VidGear v0.2.6”, 2022.  [URL](#)  [↑21](#)
- [11] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, Jia-Bin Huang. “Hybrid neural fusion for full-frame video stabilization”, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2299–2308.  [arXiv](#)  [2102.06205](#) [↑19](#)
- [12] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, Bing Zeng. “MeshFlow: minimum latency online video stabilization”, *ECCV 2016: Computer Vision – ECCV 2016 (October 11–14, 2016, Amsterdam, The Netherlands)*, Lecture Notes in Computer Science, vol. **9910**, 2016, pp. 800–815.  [URL](#)  [↑22](#)
- [13] M. Grundmann, V. Kwatra, I. Essa. “Auto-directed video stabilization with robust L1 optimal camera paths”, *CVPR ’11: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (20–25 June 2011, Colorado Springs, CO, USA)*, IEEE Computer Society, 2011, ISBN 978-1-4577-0394-2, pp. 225–232.  [↑23](#)

- [14] M. Grundmann, V. Kwatra, I. Essa. “Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths”, *CVPR ’11: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (20–25 June 2011, Colorado Springs, CO, USA), IEEE Computer Society, 2011, pp. 225–232. 
- ↑²³
- [15] Yufei Xu, Jing Zhang, Stephen J. Maybank, Dacheng Tao. “DUT: learning video stabilization by simply watching unstable videos”, *IEEE Transactions on Image Processing*, **31** (2022), pp. 4306–4320.  arXiv  2011.14574 ↑²³
- [16] Jinsoo Choi, In So Kweon. “Deep iterative frame interpolation for full-frame video stabilization”, *ACM Transactions on Graphics*, **39**:1 (2019), id. 4, 9 pp.  arXiv  1909.02641 ↑²³
- [17] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, S.-M. Hu. “Deep online video stabilization with multi-grid warping transformation learning”, *IEEE Transactions on Image Processing*, **28**:5 (2019), pp. 2283–2292.  arXiv  1909.02641 ↑²³

Received	11.03.2023;
approved after reviewing	23.03.2023;
accepted for publication	28.03.2023;
published online	07.07.2023.

Recommended by

prof. A. M. Elizarov

Information about the author:



Vitaly Petrovich Fralenko

PhD, Leading Researcher at RCMS Ailamazyan Program Systems Institute. The field of scientific interests: intellectual data analysis and images recognition, artificial intelligence and decision making, parallel-conveyor computing, network security, diagnosis of complex technical systems, graphic interfaces.



0000-0003-0123-3773

e-mail: alarmod@pereslavl.ru

The author declare no conflicts of interests.



Моделирование поведения двухуровневой квантовой системы с использованием масштабируемых регулярных сеток

Анатолий Дмитриевич **Панферов**^{1✉}, Наталия Владимировна **Поснова**²,
Анастасия Алексеевна **Ульянова**³

Саратовский государственный университет им. Н. Г. Чернышевского, Саратов, Россия

[✉]panferovad@sgu.ru

Аннотация. Представлены результаты программной реализации покрытия двумерной области допустимых состояний моделируемой двухуровневой квантовой системы регулярными сетками различного масштаба. Метод был предложен и реализован применительно к задаче воспроизведения поведения графена во внешних электрических полях с использованием квантового кинетического уравнения. Численное решение кинетического уравнения позволяет точно и полно воспроизводить поведение модели, однако проблемой является минимизация необходимых вычислительных ресурсов при обеспечении требуемого качества вычисления наблюдаемых параметров. Сложность задачи прямо пропорциональна количеству узлов сетки в двумерном пространстве состояний, в каждом из которых для воспроизведения эволюции заселенности верхнего энергетического уровня решается система обыкновенных дифференциальных уравнений. Исследование особенностей поведения модели позволило разработать и реализовать экономичную процедуру построения расчетной сетки, обеспечивающую локализацию области формирования возбуждённых состояний и точное воспроизведение наблюдаемых параметров процесса.

Ключевые слова и фразы: численное моделирование, графен, функция распределения носителей заряда, оптимальный выбор расчетной сетки

Благодарности: Исследование выполнено за счет гранта Российского научного фонда № 23-21-00047^{URL}

Для цитирования: Панферов А. Д., Поснова Н. В., Ульянова А. А. *Моделирование поведения двухуровневой квантовой системы с использованием масштабируемых регулярных сеток* // Программные системы: теория и приложения. 2023. Т. 14. № 2(57). С. 27–47. https://psta.psiras.ru/read/psta2023_2_27-47.pdf

Введение

Активный поиск новых материалов для современной электроники повлек за собой развитие исследований структур с пониженной размерностью. Технологически удобными объектами такого типа являются двумерные пленки, способные устойчиво существовать в свободном пространстве или на подложках. Свойства электронов в них и особенности процессов их взаимодействия с внешними электрическими полями могут быть описаны только с привлечением принципов и подходов квантовой механики. Наиболее известным примером материалов этого типа является графен, доступ к образцам которого позволил разработать и экспериментально верифицировать уже ставшую классической модель безмассовых фермионов [1, 2]. В настоящее время интерес к этому материалу обусловлен особенностями его отклика на действие внешних электрических полей [3–8].

Для описания процессов образования и эволюции носителей заряда в двумерном монослое графена в условиях действия внешнего классического электрического поля в работах [9–11] был предложен и развит подход, основанный на использовании квантового кинетического уравнения. Метод квантового кинетического уравнения достаточно универсален и может использоваться при рассмотрении других аналогичных материалов. В данном подходе подсистема слабосвязанных электронов материала рассматривается как квантовая система с двумя энергетическими уровнями. Конкретный вид закона дисперсии (зависимости энергии состояний от импульса) может быть различным. В простейшем случае он соответствует приближению безмассовых фермионов, но может использоваться и строгая модель взаимодействия ближайших соседей. Кинетическое уравнение определяет эволюцию функции распределения $f(\vec{p}, t)$ носителей по возбуждённым состояниям в условиях внешнего воздействия и может быть представлено в виде системы обыкновенных дифференциальных уравнений, определяемой через гамильтониан модели.

Трудности использования этого подхода обусловлены тем обстоятельством, что лежащая в его основе система уравнений может решаться только численными методами и для каждой рассматриваемой точки импульсного пространства решение необходимо искать независимо. Это приводит к высокой вычислительной сложности процедуры получения наблюдаемых характеристик. В работе [12] была продемонстрирована возможность использования адаптивной сетки на основе квадродерева для вычисления конечного состояния $f(\vec{p}, t \rightarrow \infty)$, формирующегося

после завершения действия внешнего электрического поля с импульсным характером зависимости от времени.

Использование сетки переменной плотности в области локализации максимальных значений функции распределения обеспечивает существенную экономию ресурсов по сравнению с использованием регулярных сеток. Однако проблема предварительного определения местоположения области локализации максимальных значений не имела эффективного решения. На основе практики моделирования процессов в рассматриваемых системах было обнаружено, что локализацию областей концентрации возбуждённых состояний (где требуется использовать плотные сетки) удобнее выполнять по данным о $f(\bar{p}, t)$ не после завершения внешнего воздействия, а во время его. Это было продемонстрировано в работе [13], в которой сопоставлялись особенности поведения в импульсном пространстве $f(\bar{p}, t \rightarrow \infty)$ и $f(\bar{p}, t_{E_{max}})$, где $t_{E_{max}}$ определялось как момент времени с максимальным значением внешнего воздействия (напряжённости внешнего электрического поля).

В развитии этой идеи для моделей рассматриваемого типа в представляемой работе выполнены исследования и сопоставление поведения в импульсном пространстве функции распределения в конечном состоянии $f(\bar{p}, t \rightarrow \infty)$, в момент максимума внешнего воздействия $f(\bar{p}, t_{E_{max}})$, а также максимальных значений функции распределения в выбранной точке импульсного пространства за все время действия внешнего поля $f_{max}(\bar{p})$ и средних значений функции распределения за время действия внешнего поля $f_{avg}(\bar{p})$. Это позволило разработать итерационный метод построения разномасштабных регулярных сеток, обеспечивающий в рамках унифицированной процедуры локализацию области генерации возбуждённых состояний с охватом всех рассматриваемых значений $(\bar{p})$ (полного образа первой зоны Бриллюэна) и формирование в ней плотных сеток разного масштаба.

1. Модель и её параметры

При исследовании особенностей реакции перспективных двумерных материалов на действие внешних электрических полей эффективно одноэлектронное приближение. В этом приближении квантовая эволюция электронов, не связанных жестко с конкретными атомами и имеющими возможность реагировать на воздействие внешнего поля, определяется уравнением Шрёдингера через одночастичный оператор Гамильтона, отражающий специфику внутренних полей кристаллической решетки.

В отсутствии явной зависимости гамильтониана от времени в импульсном представлении уравнение Шрёдингера принимает стационарную форму

$$(1) \quad H_{\bar{p}}\Phi(\bar{p}, t) = \epsilon(\bar{p})\Phi(\bar{p}, t),$$

определяющую спектр состояний (зонную структуру) $\epsilon(\bar{p})$ и соответствующие этим состояниям собственные функции $\Phi(\bar{p}, t)$. Эволюция стационарных состояний во времени определяется в явной форме фазовым множителем $\exp(-i\epsilon(\bar{p})t/\hbar)$.

Появление классического внешнего электрического поля выводит систему из стационарного состояния и меняет характер её эволюции. Действие классического поля на квантовую систему описывается через модификацию гамильтониана путем замены в нем оператора импульса на квазиимпульс по правилу

$$(2) \quad p_k \rightarrow P_k = p_k - eA_k(t), \quad (k = 1, 2),$$

где $A_k(t)$ — компоненты векторного потенциала внешнего поля и мы принимаем, что заряд электронов отрицателен, а его абсолютное значение совпадает с величиной элементарного заряда e . Наблюдаемая напряжённость электрического поля $E_k(t)$ определяется выражением $E_k(t) = -\dot{A}_k(t)$.

Воспроизведение эволюции в условиях явной зависимости гамильтониана от времени для реальных систем является вычислительно сложной задачей, решаемой только численными методами. При этом удобнее работать не с самой одночастичной волновой функцией, а с определяемой через неё функцией распределения $f(\bar{p}, t)$ электронов по одночастичным энергетическим состояниям. Её эволюция может быть определена через кинетическое уравнение вида:

$$(3) \quad \begin{cases} \dot{f}(\bar{p}, t) = \frac{1}{2}\lambda(\bar{p}, t)u(\bar{p}, t), \\ \dot{u}(\bar{p}, t) = \lambda(\bar{p}, t)(1 - 2f(\bar{p}, t)) - \frac{2e}{\hbar}(\bar{p}, t)v(\bar{p}, t), \\ \dot{v}(\bar{p}, t) = \frac{2e}{\hbar}(\bar{p}, t)u(\bar{p}, t), \end{cases}$$

где $u(\bar{p}, t)$ и $v(\bar{p}, t)$ вспомогательные функции (детали можно посмотреть в работе [11]). Поведение моделируемой системы определяется переменными коэффициентами $\epsilon(\bar{p}, t)$ и $\lambda(\bar{p}, t)$. Первый из них выражается через стационарный спектр $\epsilon(\bar{p})$ из уравнения (1) путем замены (2). Второй также определяется через исходный гамильтониан системы с использованием (2).

Так, при рассмотрении двумерной гексагональной решетки графена, представляющей суперпозицию двух одинаковых подрешеток, общий вид

гамильтониана имеет вид

$$(4) \quad H(\bar{p}, t) = \begin{pmatrix} 0 & B(\bar{p}, t) \\ B^*(\bar{p}, t) & 0 \end{pmatrix}$$

с двумя собственными значениями $\sqrt{BB^*}$ и $-\sqrt{BB^*}$. Соответственно, в этом случае $\epsilon(\bar{p}, t) = \sqrt{B(\bar{p}, t)B^*(\bar{p}, t)}$, а выражение для определения второго коэффициента из системы уравнений (3) имеет вид

$$(5) \quad \lambda(\bar{p}, t) = \frac{\dot{B}'(\bar{p}, t)B''(\bar{p}, t) - B'(\bar{p}, t)\dot{B}''(\bar{p}, t)}{\varepsilon^2(\bar{p}, t)},$$

где $B'(\bar{p}, t)$ и $B''(\bar{p}, t)$ – вещественная и мнимая части матричного элемента гамильтониана (4):

$$(6) \quad B(\bar{p}, t) = B'(\bar{p}, t) + iB''(\bar{p}, t).$$

В реальных случаях, например, при использовании модели взаимодействия ближайших соседей, явный вид этих выражений может быть достаточно сложным.

Периодической пространственной структуре моделируемого материала соответствует периодическая обратная решетка. Для случая графена её вид представлен на рисунке 1:

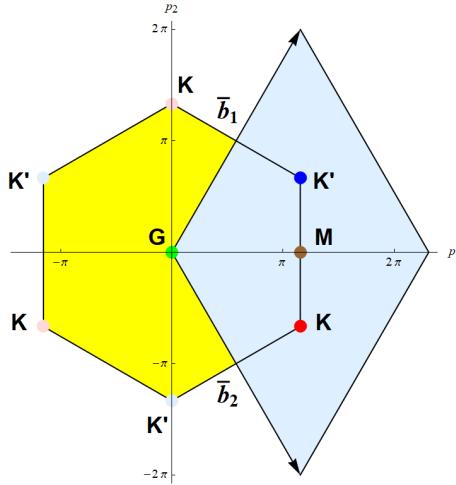


Рисунок 1. Структура обратной решетки графена. Единица измерения $[p] = \frac{\hbar}{\sqrt{3}d}$ где $d = 0.142$ нм – длина ребра шестигранной решетки в прямом пространстве.

При моделировании необходимо воспроизвести эволюцию функции распределения (решить систему уравнений (3) при всех рассматриваемых значениях $\bar{p}$) в первой зоне Бриллюэны (желтый шестиугольник) или (в силу периодичности) эквиваленте этой зоны. В качестве эквивалента была выбрана элементарная ячейка обратной решетки, построенная на векторах трансляций $\bar{b}_1$ и $\bar{b}_2$ с центром в точке M , которая будет использоваться в качестве начала системы координат. При таком выборе области определения для функции распределения в ней оказываются две неэквивалентные точки Дирака K и K' , расположенные симметрично относительно начала отсчета системы координат на оси ординат.

Для реализации возможностей модели разработан и развивается комплекс программ. Отработка его алгоритмов и функционала осуществляется поэтапно. Вначале решения тестируются на макете системы, реализованном средствами Wolfram Mathematica. Затем оформляются в виде модулей на C/C++.

Наиболее ресурсоемким является процесс решения кинетического уравнения (3), представляющего из себя систему обыкновенных дифференциальных уравнений. На этапе работы с макетом для этого используются хорошо зарекомендовавшие себя решатели Mathematica. Модули на C реализованы с использованием библиотеки GSL.

Разработка вычислительного модуля, ориентированного на использование ресурсов графических сопроцессоров NVIDIA, потребовала перехода на решатели Odeint из библиотеки Boost C++, хорошо интегрируемые в программы на CUDA. Все эти решатели могут использовать различные алгоритмы. Проведённое тестирование показало, что в случае рассматриваемого уравнения преимущество имеют явные методы высокого порядка, основанные на классе методов Рунге-Кутты: `runge_kutta_cash_karp54`, `runge_kutta_dopri5`, `runge_kutta_fehlberg78`.

Важным с точки зрения организации работы программы является возможность решать системы уравнений (3) в разных точках $\bar{p}$ независимо (параллельно). Такой параллелизм по данным без обмена промежуточными результатами позволяет эффективно использовать ресурсы современных многоядерных процессоров, графических ускорителей и мультикомпьютеров. На уровне макета используются возможности распараллеливания самой Mathematica. Решения на C/C++ используют MPI, что позволяет запускать их на мультикомпьютерах.

Наблюдаемые физические параметры $L(t)$, воспроизведение которых является целью моделирования, определяются через усреднение соответствующих величин вида

$$(7) \quad L(t) = \int l(\bar{p}, t) f(\bar{p}, t) d\bar{p}$$

по вычисленной функции распределения, где интегрирование необходимо выполнить в пределах элементарной ячейки обратной решетки. В отсутствии аналитического представления $f(\bar{p}, t)$ это интегрирование может быть выполнено только численно. Но, если зависимость $f(\bar{p}, t)$ от времени при заданном значении $\bar{p}$ определена системой уравнений (3), информации о поведении функции распределения в двумерном импульсном пространстве при фиксированном значении t модель в явной форме не содержит. По этой причине возникает задача о покрытии области определения $f(\bar{p}, t)$ в импульсном пространстве некоторой сеткой, определив на которой значения функции распределения путем решения системы уравнений (3) в каждом узле, можно получить достаточно точные оценки для наблюдаемых (7). Предложенный вариант решения этой задачи излагается далее.

2. Локализация области генерации возбуждённых состояний по характеристикам функции распределения носителей

Максимальная энергия возбуждённых состояний достигает в графене примерно 9 эВ. При этом в окрестностях точек Дирака она близка к нулю. Это приводит к крайне неоднородному распределению возбуждённых состояний. Подробно эта проблема обсуждалась в работе [13]. Там же была продемонстрирована принципиальная разница в поведении $f(\bar{p})$ в конечном состоянии после выключения поля и в его присутствии.

На основании отмеченной особенности было принято решение исследовать характер поведения в импульсном пространстве таких характеристик функции распределения, как среднее значение за время возмущающего воздействия и максимальное её значение за время действия возмущающего воздействия. Обозначим их $f_{avg}(\bar{p})$ и $f_{max}(\bar{p})$. На рисунке 2 показано в высоком разрешении поведение функции распределения в окрестностях одной из точек Дирака для короткого импульса электромагнитного поля с круговой поляризацией. Области, в которых присутствуют достаточно большие значения из интервала $[0.1 \leq f(\bar{p}) \leq 1.0]$, совпадают.

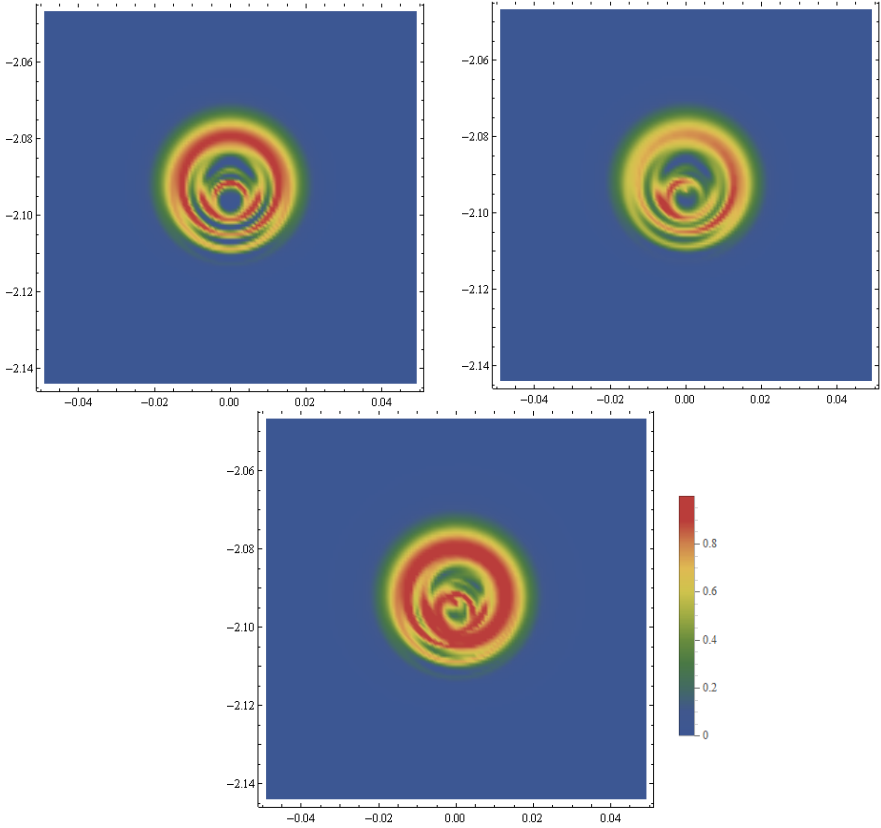


РИСУНОК 2. Функция распределения после завершения действия возмущения $f(\bar{p}, t \rightarrow \infty)$, её усреднённые за время действия возмущения значения $f_{avg}(\bar{p})$ и максимальные зарегистрированные значения $f_{max}(\bar{p})$ в окрестностях нижней точки Дирака K (см. рисунок 1).

Сравним поведение представленных функций в глобальном масштабе. На рисунке 3 показаны $f_{avg}(\bar{p})$ и $f_{max}(\bar{p})$ для отрицательных значений p_1 . В силу очень большого диапазона значений выполнено приведение к логарифмическому масштабу преобразованием вида $f \rightarrow 10 \log_{10} f$.

Ожидаемо $f_{avg}(\bar{p}) < f_{max}(\bar{p})$, но при этом их поведение во всей рассматриваемой области гладкое и очень похожее $f_{avg}(\bar{p}) \sim f_{max}(\bar{p})$. Исходя из этого подобия далее будем использовать $f_{max}(\bar{p})$ и сравним её

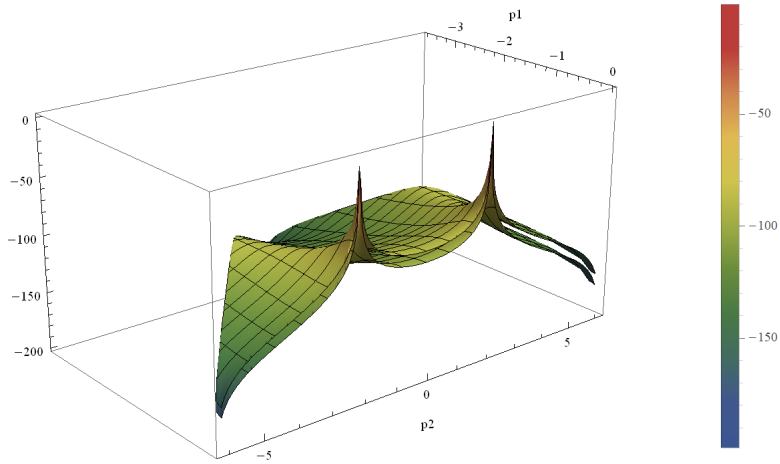


РИСУНОК 3. Глобальное поведение значений $f_{avg}(\bar{p})$ и $f_{max}(\bar{p})$.

с функцией распределения для конечного состояния модели $f(\bar{p}, t \rightarrow \infty)$. Такое сравнение приведено на рисунке 4 с аналогичным приведением к логарифмическому масштабу.

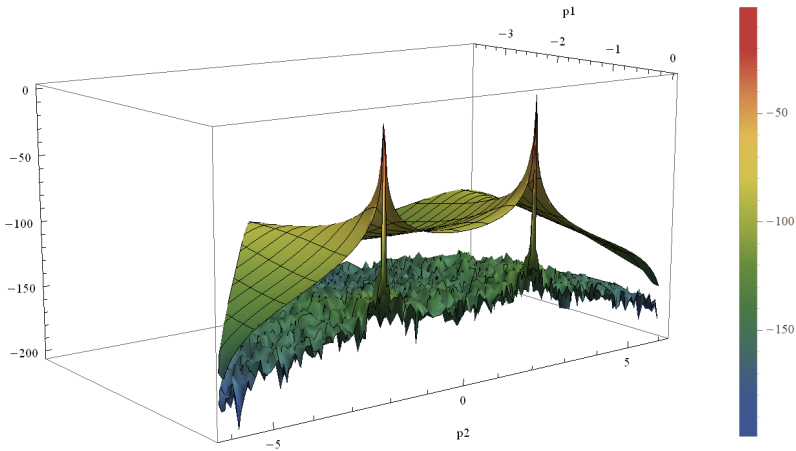


РИСУНОК 4. Глобальное поведение значений $f_{max}(\bar{p})$ и функции распределения для конечного состояния модели $f(\bar{p}, t \rightarrow \infty)$.

Представленные результаты полностью согласуются с выводами,

сделанными ранее в работе [13]. Корректное воспроизведение $f(\bar{p}, t \rightarrow \infty)$ возможно только в ограниченных областях импульсного пространства. Вне этих областей значения функции недоступны для вычисления при использовании 64-битной арифметики. При этом имеет место строгое совпадение мест локализации максимумов $f(\bar{p}, t \rightarrow \infty)$ и $f_{\max}(\bar{p})$, что позволяет использовать последнюю в качестве удобного инструмента для определения положения максимумов.

При ограничении условием $f_{\max}(\bar{p}) \lesssim 0.1$ эта функция ведет себя гладко во всей области определения. Для локализации по её значениям областей с максимумами будем использовать регулярную сетку $N \times N$, определяющую в импульсном пространстве дискретный набор точек $\{p_i, p_j\}$. На сетке будут отбираться узлы с $f_{\max}(p_i, p_j) > f_{\lim}$, где $f_{\lim}$ – нижняя граница интересующего нас диапазона значений. Уже на регулярной сетке с $N = 32$ ($\Delta p_1 = \Delta p_2 = 4\pi/32$) для использовавшихся выше параметров импульса локализация уверенно выполняется при $f_{\lim} = 1.0 \times 10^{-9}$. На рисунке 5 представлено распределение значений на такой сетке для отрицательного полупространства $p_1 \leq 0$.

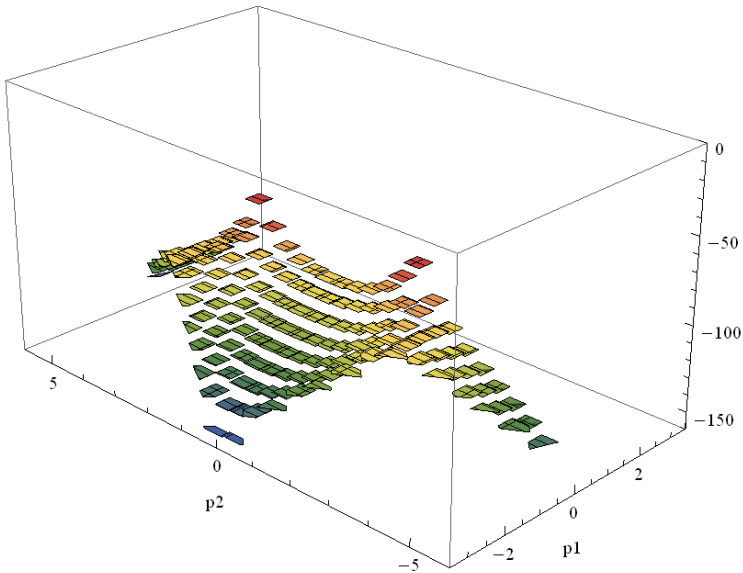


РИСУНОК 5. Распределение значений $10 \log(f_{\max}(p_i, p_j))$ при $N = 32$.

На рисунке 6 выделены только значения, удовлетворяющие условию $f_{max}(p_i, p_j) > 1.0 \times 10^{-9}$ для всей области допустимых значений (p_1, p_2) .

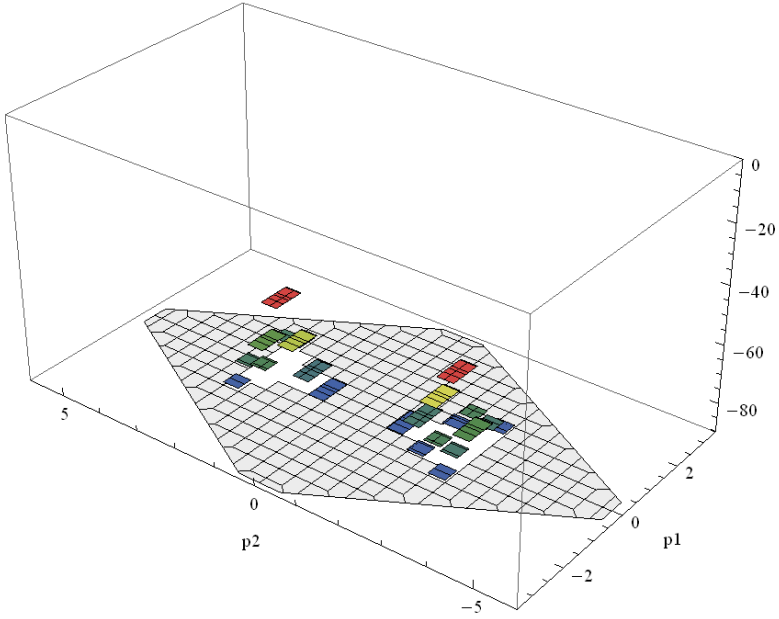


РИСУНОК 6. Выборка значений $10 \log(f_{max}(p_i, p_j))$ для $f_{max}(p_i, p_j) > 1.0 \times 10^{-9}$ при $N = 32$.

Необходимо отметить, что выполненная процедура не предоставляет точных координат области, показанной на рисунке 2. Пусть критерий попадания в эту область $0.01 \lesssim f_{max}(p_i, p_j) \lesssim 1.0$. Максимальное же значение $f_{max}(p_i, p_j)$ из полученных на сетке 32×32 и показанных на рисунках 5 и 6 всего 3.99×10^{-7} . Возможное использование на первом этапе сеток с меньшим шагом принципиально не решает эту проблему. Представленные результаты получены только для одного из многих возможных сочетаний параметров моделируемого процесса, но они вполне показательны.

3. Выборочное масштабирование сетки

Продемонстрированная выше процедура позволяет уменьшить область поиска до нескольких ячеек исходной сетки. Конкретное их число будет

определяться выбранным уровнем отсечения f_{lim} . При $f_{lim} = 1.0 \times 10^{-9}$ значение $f_{max}(p_i, p_j)$ превышает этот уровень в 34 узлах сетки. При $f_{lim} = 1.0 \times 10^{-7}$ пороговое значение будет превышено только в 4 узлах. Эти узлы расположены попарно в двух разнесенных областях.

Покажем, что этого достаточно для локализации области, где $f_{max}(\bar{p})$ имеет близкие к единице значения. Дальнейшее исследование поведения $f_{max}(\bar{p})$ в окрестностях выбранных точек выполним по аналогии с первым шагом, разбивая каждую ячейку с соответствующей узловой точкой в середине регулярной сеткой $M \times M$. Сохраним значение $M = 32$. Результаты показаны на рисунке 7. Четыре ячейки исходной сетки (далее будем говорить о ней как о сетке 0-го уровня) объединены попарно.

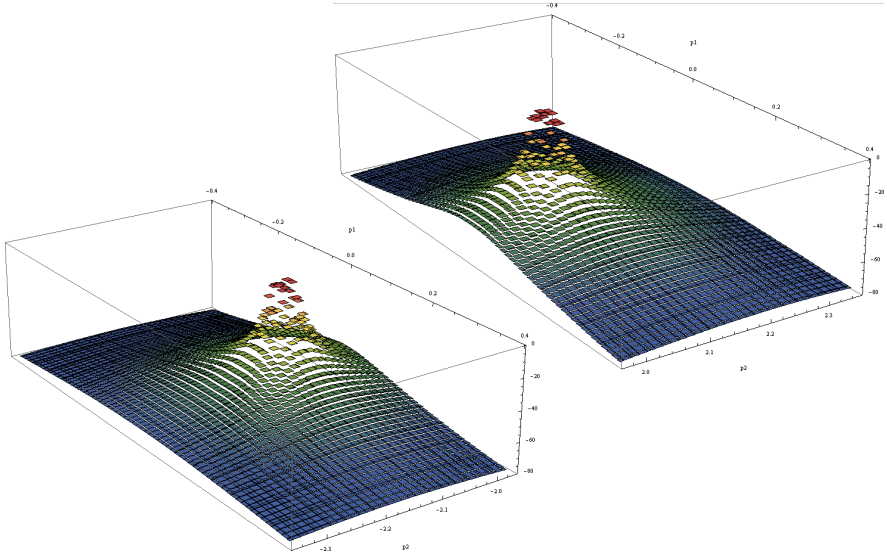


Рисунок 7. Распределение значений $10 \log(f_{max}(p_i, p_j))$ для двух пар ячеек нулевого уровня.

На рисунке 8 представлена общая картина, получающаяся при объединении новых результатов с результатами для узлов нулевого уровня:

На показанных на рисунке 7 сетках первого уровня присутствует 32 узла со значениями $f_{max}(p_i, p_j) \gtrsim 0.01$. Продолжим процедуру и выполним на них следующий шаг дискретизации с использованием

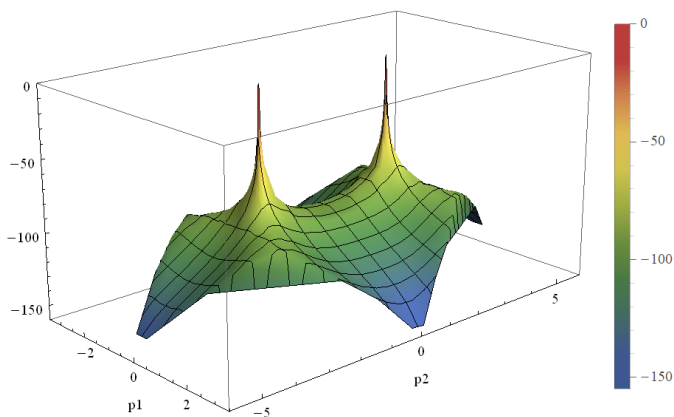


Рисунок 8. Совмещение результатов, полученных для $f_{max}(p_i, p_j)$ на сетках нулевого и первого уровней.

масштабированной регулярной сетки 32×32 . На рисунке 9 показаны результаты для ячейки первого уровня, соответствующей максимуму на правой части Рисунка 7.

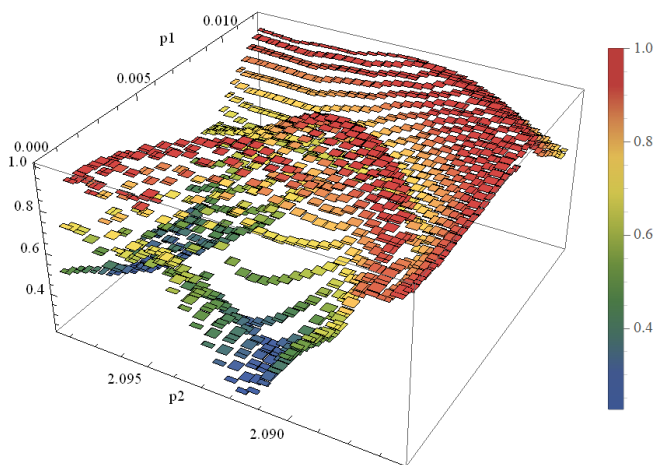


Рисунок 9. Распределение $f_{max}(p_i, p_j)$ в ячейке сетки первого уровня.

На этом этапе становится доступна мелкомасштабная структура

функции распределения, причем все представленные значения лежат в диапазоне $0.1 \lesssim f_{\max}(p_i, p_j) \lesssim 1.0$.

Полная картина для одного из максимумов, показанных на рисунках 7 и 8, сгруппирована из данных для 16 ячеек сетки первого уровня на рисунке 10. На нем представлены значения $0.01 \lesssim f_{\max}(p_i, p_j) \lesssim 1.0$,

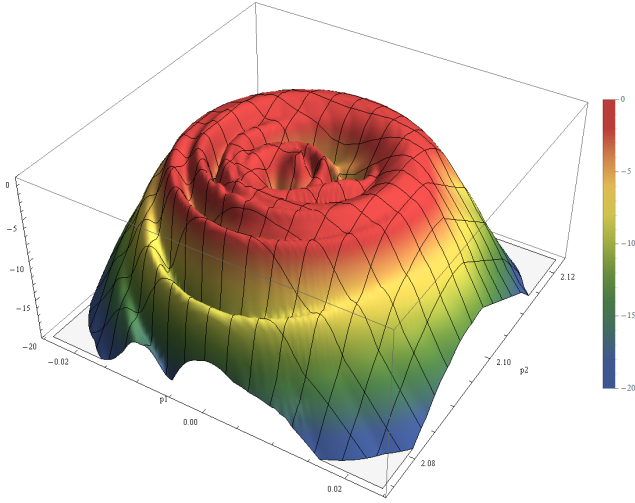


Рисунок 10. Распределение значений $10 \log(f_{\max}(p_i, p_j))$ при $N = 32$.

приведённые к логарифмическому масштабу.

В результате двух итераций получена гибридная сетка с тремя масштабами покрытия исследуемой области с отношением размеров ячеек $1 : 1/32 : 1/32/32$. Общее количество узлов гибридной сетки $\simeq 3.7 \times 10^4$. Процедура обеспечивает как локализацию областей с максимальными значениями функции распределения, определяющими наблюдаемые параметры моделируемого процесса, так и точное воспроизведение мелко-масштабной структуры функции распределения в двух неэквивалентных областях.

Располагая такой сеткой, можно экономично и точно исследовать поведение модели на всем протяжении процесса её взаимодействия с импульсом внешнего поля. На рисунке 11 показаны функции распределения для ряда последовательных моментов времени $\{t_0, t_1, t_2, t_3\}$. Первый из них соответствует максимуму напряженности внешнего электрического

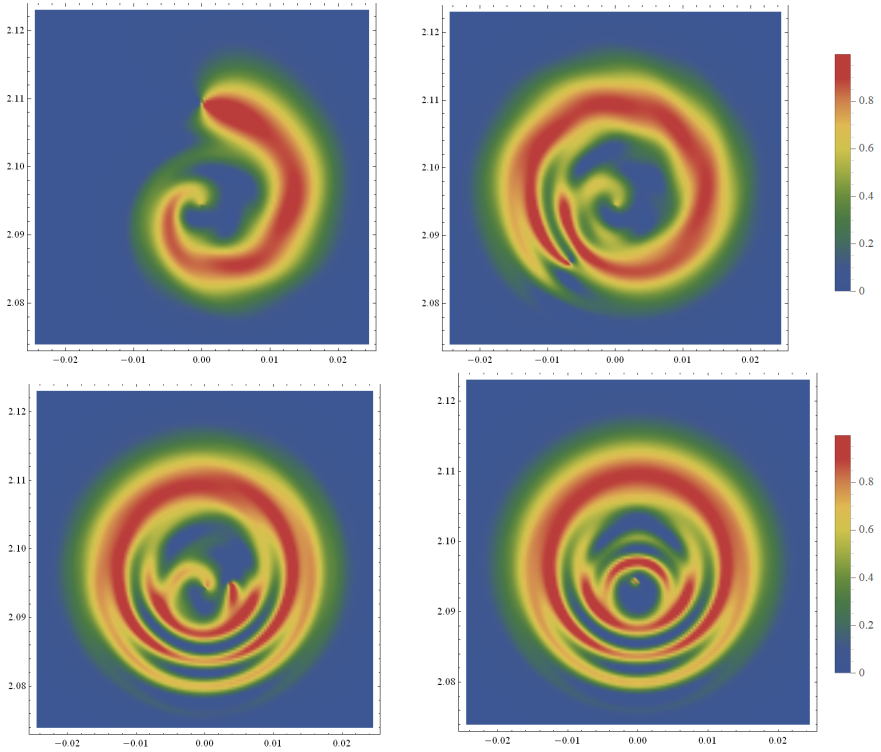


Рисунок 11. Воспроизведение поведения $f(p_1, p_2, t)$ на примере четырех последовательных моментов времени.

поля, последний — полному выключению внешнего воздействия, $\{t_1, t_2\}$ — промежуточные значения. Приведены результаты для области покрытия сеткой второго уровня, показанной на рисунке 10. Во всех случаях обеспечивается воспроизведение тонкой структуры распределения возбуждённых состояний и полный охват области их локализации.

Оценку эффективности предложенной процедуры можно сделать сравнивая построенную сетку переменного масштаба с эквивалентной ей по обеспечиваемой детализации равномерной сеткой. В области воспроизведения тонкой структуры функции распределения в окрестностях её максимумов (например, рисунок 9) расстояние между узлами построенной сетки составляет $4\pi/32/32/32$. Если воспользоваться равномерной сеткой с таким шагом, для покрытия всей области определения функции распределения в импульсном пространстве потребуется $\simeq 3.1 \times 10^8$ узлов, что почти на четыре порядка больше числа узлов в построенной

сетке переменного масштаба. Обеспечиваемая экономия вычислительных ресурсов делает возможным моделирование процессов рассматриваемого типа на современных высокопроизводительных рабочих станциях.

Фактические затраты времени на каждый узел сетки зависят от параметров моделируемого процесса, программной реализации, используемой аппаратной платформы и ряда других факторов. Демонстрация процедуры построения сетки выполнена при фиксированных параметрах моделируемого процесса на системе с CPU Intel(R) Core(TM) i5-11400 с 6 физическими ядрами и базовой тактовой частотой 2.60 GHz, Кэш L3 12.0 MB, объём памяти 32.0 GB. Все шаги процедуры, включая решение системы уравнений (3), были реализованы с использованием возможностей Mathematica. Характерные затраты времени на воспроизведение поведения модели в каждом узле строящейся сетки оцениваются в 0.1239 с при вычислениях в последовательном режиме (используется одно ядро) или 0.02065 с при параллельной работе на 6 физических ядрах. Соответственно, оценка времени на воспроизведение поведения модели с использованием всей построенной сетки составляет 764.0 с при полном использовании ресурсов системы.

Заключение

Численное моделирование многочастичных систем, поведение которых в определяющей мере подчиняется законам квантовой механики, ресурсоемко и требует оптимизации всех элементов процедуры. В представленной работе продемонстрирована реализация метода построения расчетной сетки для покрытия области определения функции распределения носителей заряда в двумерном импульсном пространстве модельной системы с двумя энергетическими уровнями. Такая модель позволяет описывать поведение перспективных двумерных материалов в условиях действия на них внешних электрических полей. Например, графена, причем как в наиболее распространенном упрощении приближения безмассовых фермионов, так и в более строгой модели взаимодействующих ближайших соседей. Предложенный метод позволяет в рамках одной процедуры совместить поиск области локализации возбуждений в импульсном пространстве модели и воспроизведение детальной картины их распределения в импульсном пространстве и во времени.

Список литературы

- [1] Novoselov K. S., Geim A. K., Morozov S. V., Jiang D., Katsnelson M. I., Grigorieva I. V., Dubonos S. V., Firsov A. A. *Two-dimensional gas of massless dirac fermions in graphene* // Nature.— 2005.— Vol. **438**.— Pp. 197–200.  [↑28](#)
- [2] Castro Neto A. H., Guinea F., Peres N. M. R., Novoselov K. S., Geim A. K. *The electronic properties of graphene* // Rev. Mod. Phys.— 2009.— Vol. **81**.— No. 1.— id. 109.  [↑28](#)

- [3] Glazov M. M., Ganichev S. D. *High frequency electric field induced nonlinear effects in graphene* // Physics Reports.– 2014.– Vol. **535**.– No. 3.– Pp. 101–138. ^{↑28}
- [4] Bowlan P., Martinez-Moreno E., Reimann K., Elsaesser T., Woerner M. *Ultrafast terahertz response of multilayer graphene in the nonperturbative regime* // Phys. Rev. B.– 2014.– Vol. **89**.– id. 041408. ^{↑28}
- [5] Baudisch M., Marini A., Cox J. D., Zhu T., Silva F., Teichmann S., Massicotte M., Koppens F., Levitov L. S., García de Abajo F. J., Biegert J. *Ultrafast nonlinear optical response of Dirac fermions in graphene* // Nature Communications.– 2018.– Vol. **9**.– id. 1018.– 6 pp. ^{↑28}
- [6] Zi-Yu Chen, Rui Gin *Circularly polarized extreme ultraviolet high harmonic generation in graphene* // Optics Express.– 2019.– Vol. **27**.– No. 3.– Pp. 3761–3770. ^{↑28}
- [7] Wenwen Mao, Angel Rubio, Shunsuke A. Sato *Terahertz-induced high-order harmonic generation and nonlinear charge transport in graphene* // Phys. Rev. B.– 2022.– Vol. **106**.– id. 024313. ^{↑28}
- [8] Soonyoung Cha, Minjeong Kim, Youngjae Kim, Shinyoung Choi, Sejong Kang, Hoon Kim, Sangho Yoon, Gunho Moon, Taeho Kim, Ye Won Lee, Gil Young Cho, Moon Jeong Park, Cheol-Joo Kim, Kim B. J., JaeDong Lee, Moon-Ho Jo, Jonghwan Kim *Gate-tunable quantum pathways of high harmonic generation in graphene* // Nature Communications.– 2022.– Vol. **13**.– id. 6630.– 10 pp. ^{↑28}
- [9] Smolyansky S. A., Churochkin D. V., Dmitriev V. V., Panferov A. D., Kämpfer B. *Residual currents generated from vacuum by an electric field pulse in 2+1 dimensional QED models* // EPJ Web of Conferences.– 2017.– Vol. **138**.– id. 06004.– 5 pp. ^{↑28}
- [10] Panferov A. D., Smolyansky S. A., Blaschke D. B., N. Gevorgyan T. *Comparing two different descriptions of the I-V characteristic of graphene: theory and experiment* // EPJ Web of Conferences.– 2019.– Vol. **204**.– id. 06008.– 6 pp. ^{↑28}
- [11] Smolyansky S. A., Panferov A. D., Blaschke D. B., Gevorgyan N. T. *Nonperturbative kinetic description of electron-hole excitations in graphene in a time dependent electric field of arbitrary polarization* // Particles.– 2019.– Vol. **2**.– No. 2.– Pp. 208–230. ^{↑28, 30}
- [12] Панферов А. Д., Маханьков А. В., Трунов А. А. *Использование адаптивной сетки на основе квадродерева для моделирования конечного состояния квантово-полевой системы при импульсном внешнем воздействии* // Программные системы: теория и приложения.– 2020.– Т. **11**.– № 1.– С. 79–92. ^{↑28}
- [13] Панферов А. Д., Новиков Н. А., Трунов А. А. *Моделирование поведения графена во внешних электрических полях* // Программные системы: теория и приложения.– 2021.– Т. **12**.– С. 3. ^{↑29, 33, 36}

Поступила в редакцию
одобрена после рецензирования
принята к публикации
опубликована онлайн

20.02.2023;
12.05.2023;
04.06.2023;
07.07.2023.

Рекомендовал к публикации

д.т.н. А. М. Цирлин

Информация об авторах:



Анатолий Дмитриевич Панферов

К.ф.-м.н., зам. начальника УЦИТ Саратовского государственного университета им. Н. Г. Чернышевского по научно-производственной деятельности. Научные интересы: высокопроизводительные вычисления, параллельное программирование, численное решение квантовых кинетических уравнений, моделирование процессов вакуумного рождение частиц в КЭД, генерации носителей в полупроводниках в том числе бесщелевых, процессов на ранних стадиях столкновения релятивистских ядер



0000-0003-2332-0982

e-mail: panferovad@sgu.ru



Наталья Владимировна Поснова

Магистрант Саратовского государственного университета им. Н. Г. Чернышевского. Научные интересы: моделирование физических процессов на высокопроизводительных вычислительных системах, параллельное программирование.



0000-0003-1259-1867

e-mail: natalie.posnova@gmail.com



Анастасия Алексеевна Ульянова

Магистрант Саратовского государственного университета им. Н. Г. Чернышевского. Научные интересы: моделирование физических процессов на высокопроизводительных вычислительных системах, параллельное программирование.



0000-0001-9519-9822

e-mail: ulyanova.nastiya@yandex.ru

Вклад авторов: *А. Д. Панферов* – 99.95% (идея, методология, создание стилевого файла, сбор материала, написание черновой версии, доработка и редактирование, визуализация, администрирование); *Н. В. Поснова* – 0.05% (доработка и редактирование); *А. А. Ульянова* – 0.05% (доработка и редактирование).

Авторы заявляют об отсутствии конфликта интересов.



Simulation the behavior of a two-level quantum system using scalable regular grids

Anatolii Dmitrievich **Panferov**¹, Natalia Vladimirovna **Posnova**²,
Anastasia Alekseevna **Ulyanova**³

Saratov State University, Saratov, Russia

¹panferovad@sgu.ru

Abstract. The paper demonstrates a method for constructing a grid with a changing scale for modeling the behavior of a two-level quantum system. The method was proposed and implemented in relation to the problem of reproducing the behavior of graphene in external electric fields using the quantum kinetic equation. A direct numerical solution of the kinetic equation makes it possible to accurately and completely reproduce the behavior of models of this type. But a large amount of computation is required for a good quality reproduction of the observed parameters. The complexity of the problem is directly proportional to the number of grid nodes in the two-dimensional space of states, in each of which a system of ordinary differential equations is solved to reproduce the evolution of the population of the upper energy level. The study of the behavior of the model made it possible to develop and implement an economical procedure for constructing a computational grid, which ensures the localization of the region of formation of excited states and exact reproduction of the observed process parameters. (*In Russian*).

Key words and phrases: numerical simulation, graphene, charge carrier distribution function, optimal choice of computational grid

2020 *Mathematics Subject Classification:* 65M50; 81T10, 81T40

Acknowledgments: The study was supported by a grant from the Russian Science Foundation No. 23-21-00047^{URL}

For citation: Anatolii D. Panferov, Natalia V. Posnova, Anastasia A. Ulyanova. *Simulation the behavior of a two-level quantum system using scalable regular grids*. Program Systems: Theory and Applications, 2023, **14**:2(57), pp. 27–47. (*In Russ.*). https://psta.psisras.ru/read/psta2023_2_27-47.pdf

References

- [1] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, A. A. Firsov. “Two-dimensional gas of massless dirac fermions in graphene”, *Nature*, **438** (2005), pp. 197–200. 
- [2] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, A. K. Geim. “The electronic properties of graphene”, *Rev. Mod. Phys.*, **81**:1 (2009), id. 109. 
- [3] M. M. Glazov, S. D. Ganichev. “High frequency electric field induced nonlinear effects in graphene”, *Physics Reports*, **535**:3 (2014), pp. 101–138. 
- [4] P. Bownan, E. Martinez-Moreno, K. Reimann, T. Elsaesser, M. Woerner. “Ultrafast terahertz response of multilayer graphene in the nonperturbative regime”, *Phys. Rev. B*, **89** (2014), id. 041408. 
- [5] M. Baudisch, A. Marini, J. D. Cox, T. Zhu, F. Silva, S. Teichmann, M. Massicotte, F. Koppens, L. S. Levitov, de Abajo F. J. García, J. Biegert. “Ultrafast nonlinear optical response of Dirac fermions in graphene”, *Nature Communications*, **9** (2018), id. 1018, 6 pp. 
- [6] Chen Zi-Yu, Gin Rui. “Circularly polarized extreme ultraviolet high harmonic generation in graphene”, *Optics Express*, **27**:3 (2019), pp. 3761–3770. 
- [7] Mao Wenwen, Rubio Angel, A. Sato Shunsuke. “Terahertz-induced high-order harmonic generation and nonlinear charge transport in graphene”, *Phys. Rev. B*, **106** (2022), id. 024313. 
- [8] Cha Soonyoung, Kim Minjeong, Kim Youngjae, Choi Shinyoung, Kang Sejong, Kim Hoon, Yoon Sangho, Moon Gunho, Kim Taeho, Won Lee Ye, Young Cho Gil, Jeong Park Moon, Kim Cheol-Joo, B. J. Kim, Lee JaeDong, Jo Moon-Ho, Kim Jonghwan. “Gate-tunable quantum pathways of high harmonic generation in graphene”, *Nature Communications*, **13** (2022), id. 6630, 10 pp. 
- [9] S. A. Smolyansky, D. V. Churochkin, V. V. Dmitriev, A. D. Panferov, B. Kämpfer. “Residual currents generated from vacuum by an electric field pulse in 2+1 dimensional QED models”, *EPJ Web of Conferences*, **138** (2017), id. 06004, 5 pp. 
- [10] A. D. Panferov, S. A. Smolyansky, D. B. Blaschke, N. T. Gevorgyan. “Comparing two different descriptions of the I-V characteristic of graphene: theory and experiment”, *EPJ Web of Conferences*, **204** (2019), id. 06008, 6 pp. 
- [11] S. A. Smolyansky, A. D. Panferov, D. B. Blaschke, N. T. Gevorgyan. “Nonperturbative kinetic description of electron-hole excitations in graphene in a time dependent electric field of arbitrary polarization”, *Particles*, **2**:2 (2019), pp. 208–230. 

- [12] A. D. Panferov, A. V. Makhankov, A. A. Trunov. “The use of an adaptive mesh based on a quadtree for modeling the final state of a quantum field system under pulsed external action”, *Program Systems: Theory and Applications*, **11**:1 (2020), pp. 93–105.  
- [13] A. D. Panferov, N. A. Novikov, A. A. Trunov. “Simulate the behavior of graphene in external electric fields”, *Program Systems: Theory and Applications*, **12** (2021), pp. 3 (in Russian).  



Ноябрь 2022: состояние и перспективы развития суперкомпьютерной отрасли в мире и в России

Николай Сергеевич **Абрамов**^{1✉}, Сергей Михайлович **Абрамов**²

Институт программных систем им. А. К. Айламазяна РАН, Вельково, Россия

^{1✉}*n-say@nsa.pereslavl.ru*

Аннотация. Статья посвящена анализу современного состояния суперкомпьютерной отрасли в мире и в России и перспектив ее развития. По методическому подходу и инструментарию данная статья является продолжением серии подобных аналитических обзоров авторов. Новизна данной работы не только в актуализации сведений с учетом последних редакций списка Top500, но и в акценте на следующее: мировая суперкомпьютерная отрасль переживает радикальную перестройку – переход от «петафлопсной эпохи» к «эксафлопсной эпохе». России с некоторым запаздыванием предстоит сделать такой же переход. Поэтому важно понимать технологические тренды и особенности решений для самых мощных систем в мире за последние годы. На это обстоятельство в работе обращается особое внимание и такой акцент при анализе трендов является одной из побудительных целей данной работы.

Ключевые слова и фразы: Top500, суперкомпьютеры, гибридные архитектуры, интерконнект

Для цитирования: Абрамов Н. С., Абрамов С. М. *Ноябрь 2022: состояние и перспективы развития суперкомпьютерной отрасли в мире и в России* // Программные системы: теория и приложения. 2023. Т. 14. № 2(57). С. 49–93.
https://psta.psir.ru/read/psta2023_2_49-93.pdf

Введение

В ведущих странах мира суперкомпьютерные технологии (СКТ) властью и обществом достаточно давно рассматриваются как единственный инструмент обеспечения конкурентных преимуществ [1]. Сегодня, в эпоху разворачивания в мире цифровой экономики, усиливаются роли суперкомпьютерной отрасли и суперкомпьютерной киберинфраструктуры (СК-инфраструктуры) ведущих стран. Учитывая планы России совершить рывок и построить в стране цифровую экономику, важными и актуальными становятся истинная оценка сегодняшнего состояния СК-инфраструктуры нашей страны, достоверное сравнение наших позиций в суперкомпьютерной отрасли с позициями ведущих стран мира, точное представление о современных тенденциях развития СКТ.

В настоящее время мировая суперкомпьютерная отрасль переживает радикальную перестройку: идет переход от «петафлопсной эпохи» к «эксафлопсной эпохе». И это связано не просто с увеличением производительности суперкомпьютеров в 1 000 раз. В процессе этого перехода пришлось искать решение большого количества проблем как в области аппаратных средств, так и в части программного обеспечения [2].

России с некоторым запаздыванием предстоит сделать такой же переход. Поэтому очень важно понимать технологические тренды и особенности решений для самых мощных систем в мире за последние годы. Это даст возможность выделить устоявшиеся решения, обеспечившие переход к эксаскейльным системам. На это обстоятельство в работе обращается особое внимание и такой акцент при анализе трендов является одной из побудительных целей данной работы.

По методическому подходу и инструментарию данная статья является продолжением серии подобных аналитических обзоров авторов [3, 4]. Новизна материала данной работы не только в актуализации сведений с учетом последних редакций списка Top500, но и в акценте на отмеченный выше переход к эксафлопсной эпохе.

В следующих разделах анализируются:

- Список Top500 – раздел 1;
- Темпы развития суперкомпьютерных технологий – раздел 2;
- Самые мощные публичные супер-ЭВМ с июня 2011 года по ноябрь 2022 года – раздел 3;
- Тенденции развития гибридных архитектур – раздел 4;
- Тенденции в технологиях интерконнекта – раздел 5;
- Тенденции в технологиях центральных процессоров – раздел 6;

- Компании-производители суперкомпьютеров – раздел 7;
- Области применения суперкомпьютеров – раздел 8;
- Состояние суперкомпьютерной инфраструктуры в мире и в России – раздел 9.

1. Список Top500

В настоящее время обсуждение «эксафлопсной эпохи» и анализ всемирного рейтинга пятисот самых производительных суперкомпьютеров в мире – списка *Top500*^{[URL](#)} активно ведется учеными. Так, например, в работе [5] показано, что скачок производительности (Flops) в 1000 раз произошел, но при этом пропускная способность памяти увеличилась незначительно. Авторы отмечают, что благодаря, в том числе, решениям с многокристальной компоновкой, лучшими сетями и охлаждением, был достигнут рубеж в 1 EFlops. Работа [6] посвящена исследованию роли Китая в суперкомпьютерной отрасли. Авторы ставят акцент на том, что кроме списка Top500 необходимо учитывать награды за достижения в области высокопроизводительных вычислений – премия *Гордона Белла*^{[URL](#)}. При анализе списка Top500 ученым выявлены три основные тенденции в развитии отрасли: широкое распространение гетерогенных архитектур при построении суперкомпьютерных систем; приложения искусственного интеллекта становятся одними из основных, применяемых в HPC-системах; повышение сложности комплексного моделирования различных процессов. Анализ системного баланса и архитектурных тенденций при построении самых мощных суперкомпьютеров приведен в [7]. Рассматриваются все доступные редакции списка Top500 (до 2020 года), исследуются тренды корреляции между производительностью и системными атрибутами (память, ввод-вывод, интерконнект, питание и др.). Так, авторы приводят энергоэффективность вычислений (отношение вычислительной мощности к потребляемой электроэнергии) и отмечают рост корреляции Пирсона между вычислительной мощностью суперкомпьютеров на высоких позициях списка Top500 и их энергоэффективностью.

Основным источником сведений для анализа в настоящей статье также является список Top500. Редакции рейтинга обновляются два раза в год, в июне и в ноябре. Таким образом с июня 1993 года (первая редакция) по ноябрь 2022 года вышло 60 редакций. В списке Top500 системы ранжируются по реальной производительности, измеренной на тесте Linpack – Linpack-производительность. Везде далее мы будем под производительностью понимать только Linpack-производительность, даже когда это явно не подчеркивается. В списке Top500 ее обозначают *RMax*.

Анализ списка проводится в долях *RMax*, методология анализа состояния суперкомпьютерной отрасли будет описана ниже.

Имея ввиду тот факт, что некоторые системы по разным причинам не представлены в списке, под определением термина *суперкомпьютер* здесь будем понимать следующее: *суперкомпьютер – это вычислительная система, имеющая такую производительность, которая соответствует производительностям машин, включенных в соответствующую редакцию списка Top500.*

При таком определении данные, доступные на сайте Top500, являются достоверным и многоплановым описанием состояния суперкомпьютерной отрасли. Причем это описание ретроспективное – доступны все шестьдесят редакций рейтинга, с шагом в полгода и весьма подробное. Каждая редакция рейтинга – это таблица из 500 записей, в каждой записи около 40 полей с различными сведениями про каждый суперкомпьютер. Ключевые моменты методики анализа этого массива данных:

- Суперкомпьютеры, входящие в список Top500, очень сильно отличаются друг от друга по своей ключевой характеристике – производительности. И, как следствие, они сильно отличаются по используемым техническим решениям, цене, потребительским свойствам и т. д. Крайне сильное «расслоение» в мире суперкомпьютеров на примере редакции Top500 за ноябрь 2022 года иллюстрирует рисунок 1.

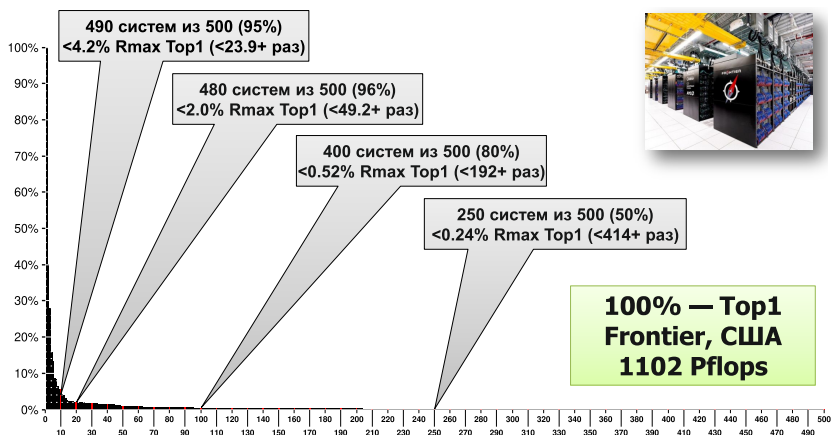


Рисунок 1. Относительная производительность 500 систем, вошедших в рейтинг Top500 в ноябре 2022 года. За 100% принята производительность системы Top1

- Как следствие, ошибочно считать суперкомпьютеры «штуками», необходимо опираться на ключевую характеристику – производительность. Например, сравнивая оснащенность различных стран суперкомпьютерной техникой, необходимо обращать больше внимания на суммарную производительность доступных стране суперкомпьютеров, а не на их количество.

2. Темпы развития суперкомпьютерных технологий

Общую оценку темпов развития СКТ можно получить из анализа достигнутой производительности суперкомпьютеров в различные моменты времени. На рисунке 2 приведена соответствующая информация.

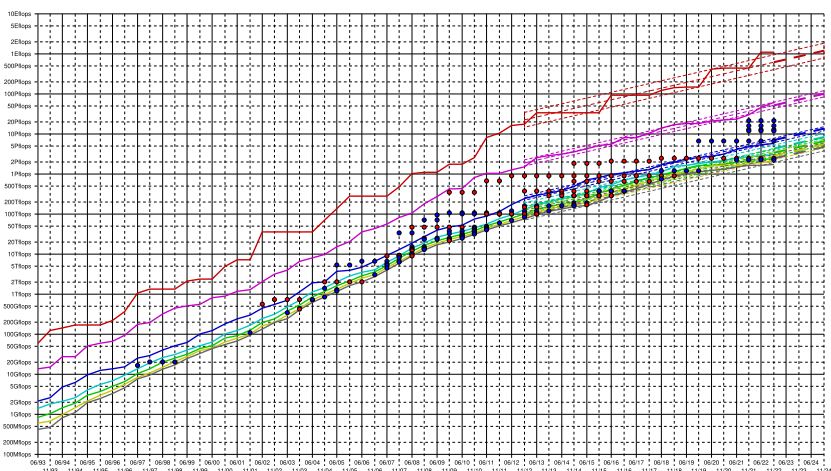


Рисунок 2. Производительность суперкомпьютеров, входящих в различные редакции рейтинга Top500

Дадим необходимые пояснения к рисунку 2:

- На оси абсцисс отмечены моменты времени (месяц и год) выхода редакций Top500, на оси ординат – производительность, шкала полулогарифмическая;
- Сплошными цветными линиями показаны графики производительности систем, занимающих различные места в рейтинге: верхний красный график – производительность системы, занимающей первое место; ниже по порядку идут графики производительности

систем, занимающих 10 место (сиреневый), 100 место (синий), 200 место (светло-бирюзовый), 300 место (светло-зеленый), 400 место (светло-травянистый) и 500 место (темно-зеленый);

- На интервале времени с июня 2012 по ноябрь 2024 года пунктирными линиями показан прогноз развития производительности систем, соответствующих 1, 10, 100, 200, 300, 400 и 500 месту в рейтинге — обозначено среднее значение прогноза и интервал 90% достоверности. Для вычисления прогноза использована групповая линейная регрессия (определяется ниже, Раздел 2.1) на отрезке последних 10 лет — с ноября 2012 по ноябрь 2022;
- Цветные точки на графике отмечают системы, установленные в России и попавшие в соответствующую редакцию списка Top500. Точки синего цвета — суперкомпьютеры, закупленные за рубежом, точки красного цвета — отечественные суперкомпьютеры.

Графики, подобные приведенным на рисунке 2, часто публикуются и служат иллюстрацией экспоненциальной скорости развития суперкомпьютерной отрасли. И правда, на первый взгляд цветные линии на рисунке 2 близки к прямым линиям, что с учетом полулогарифмической шкалы соответствует экспоненциальной зависимости. . . Однако, более аккуратный анализ показывает, что более всего графики похожи на ломаную линию из двух (как минимум) прямых отрезков — до 2008 года линейная регрессия имеет один наклон, а после 2008 года другой, более пологий.

2.1. Групповая линейная регрессия

Опишем методику упомянутого выше анализа. Занумеруем все выпуски рейтинга Top500 номерами $i \in [1, 60]$. Обозначим производительность системы, занявшей некое место $j \in [1, 500]$ в выпуске Top500 с номером i через $r_{i,j}$. Зададимся неким набором мест в рейтинге Top500: $J \subseteq [1, 500]$ и неким номером рейтинга $k \in [8, 53]$ — мы здесь отступили по 7 выпусков рейтинга слева и справа в отрезке $[1, 60]$.

Тогда анализом тренда изменения производительности методом групповой линейной регрессии для заданных k , J назовем задачу поиска такого набора параметров a и b_j , где $j \in J$, что серии данных $s_j = \{(i, \ln r_{i,j}) \mid i \in [k-7, k+7]\}$ для $j \in J$, наилучшим образом приближены линейными функциями $f_j(i) = a \cdot i + b_j$ от переменной i . А именно,

минимизируется сумма квадратов:

$$\sum_{i \in [k-7, k+7], j \in J} (a \cdot i + b_j - \ln r_{i,j})^2 \rightarrow \min$$

Задача минимизации решается методом наименьших квадратов (МНК) и, по сути, это несложная модификация стандартного алгоритма линейной регрессии: регрессия проводится для группы наборов данных s_j , заданных индексами $j \in J$, считаем, что угол наклона всех аппроксимирующих прямых линий (параметр a) одинаков, а сдвиги по оси Y (параметры b_j) различные.

По смыслу решенной задачи можем сказать, что в момент выпуска Top500 номер k , за полгода производительность систем, занимающих места $j \in J$ в среднем, возрастала в e^a раз. Значит в 1000 раз эта производительность вырастет (в среднем) за $y(k, J) = \frac{\ln 1000}{2 \cdot a}$ лет.

2.2. За сколько лет производительность суперкомпьютеров вырастала в 1000 раз?

Величина $y(k, J)$ была рассчитана для нескольких наборов J и для всех $k \in [8, 52]$. Рисунок 3 показывает результаты расчётов. На оси X вместо величины k подписана календарная дата выпуска рейтинга Top500 с номером k . По оси Y указана величина $y(k, J)$ (годы).

Видно, что на интервале 1998–2008 действительно можно было говорить, что $R_{\max}$ увеличивался в 1000 раз примерно за 11 лет. И это было верно и для лидирующих мест в суперкомпьютерном рейтинге (красная и сиреневая линия графика), и для всего рейтинга в целом (зеленая линия). Далее (2009–2016) мы наблюдаем признаки явных технических сложностей, которые сдерживали рост технологий в отрасли. В последнее время ситуация стала исправляться для старших систем в рейтинге, и сохранились негативные тенденции для рейтинга в целом – то есть, *нарастает разрыв* между старшими системами и всеми остальными.

2.3. Основные выводы из анализа темпов развития суперкомпьютерных технологий

Можем констатировать, что на рубеже 2008 года, при сохранении экспоненциального характера роста производительности, уменьшилась скорость роста – уменьшилось основание этой самой экспоненты. До 2008

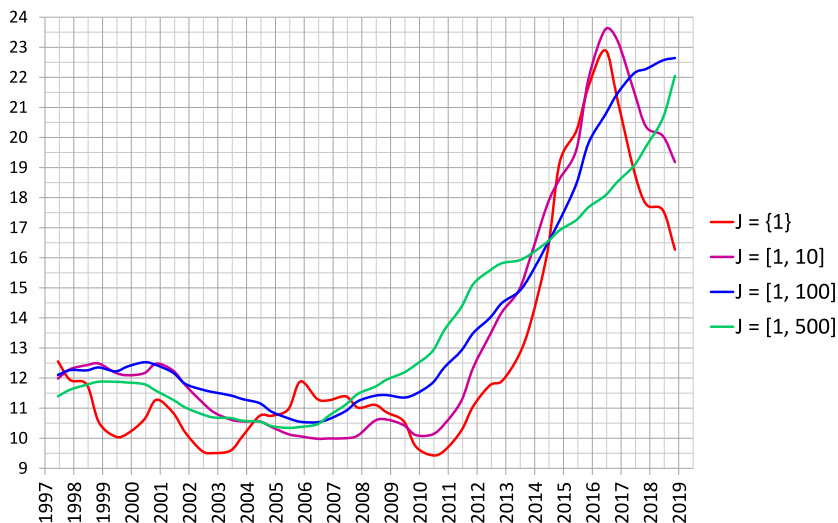


РИСУНОК 3. За сколько лет $R_{\max}$ суперкомпьютеров, занимающих места J , вырастет в среднем в 1,000 раз?

года достаточно точно выдерживалось следующее: за 18 месяцев удвоение производительности самой мощной системы в мире; за 11 лет рост ее производительности в 1000 раз. Так, рубежи производительности были преодолены: 1 Mflops (10^6 flops) в 1975 году, 1 Gflops (10^9 flops) в 1986 году, 1 Tflops (10^{12} flops) в 1997 году, 1 Pflops (10^{15} flops) в 2008 году. Если бы эта тенденция сохранилась бы, то производительность 1 Eflops (10^{18} flops) была бы достигнута в 2019 году. Но это произошло в июне 2022 года, см. рисунок 2.

Начиная с 2008 года явно изменилась скорость роста достигнутой максимальной производительности: вместо прироста «в 1000 раз за 11 лет» мы имеем прирост «в 1000 раз за 13–17 лет».

Мировая суперкомпьютерная отрасль на рубеже 2008 года, несомненно, столкнулась с научно-техническими сложностями на пути развития СКТ. Это привело к пересмотру будущих прогнозов достижения новых рубежей производительности: 1 Eflops (10^{18} flops) был достигнут в июне 2022 года, а 1 Zflops (10^{21} flops) стоит ожидать в 2035–2039 году.

3. Самые мощные публичные супер-ЭВМ с июня 2011 года по ноябрь 2022 года

В таблице 1 собраны краткие сведения о самых мощных суперкомпьютерах в мире с июня 2011 года по ноябрь 2022 года.

Таблица 1. Краткие сведения о самых мощных суперкомпьютерах в мире с июня 2011 года по ноябрь 2022 года

Редакция Top500, система, ▷ вычислительная подсистема	Rmax, электропотребление	Интерконнект. Технология охлаждения
6/2011 , K computer, Япония, Fujitsu, RIKEN AICS [8] ▷ ≈0.7М ядер , 88,128 CPU SPARC64 VIIIfx 8C 2.00 GHz 8-ядер, 4 процессора в узле	10.5 Pflops, 12.6 МВт	Собственный интерконнект. Водяное охлаждение
06/2012 , Sequoia, США, BlueGene/Q, IBM [9] ▷ ≈1.6М ядер , ≈99K CPU Power BQC 16C, 64-разрядный RISC, 1.6 ГГц 16 ядер	16.3 Pflops, 7.9 МВт	Собственный интерконнект. Водяное охлаждение
11/2012 , Titan, США, Cray XK7 [10] ▷ ≈0.56М ядер , AMDOpteron – ≈300K ядер, NVIDIA K20x – ≈260K ядер	17.6 Pflops, 8.2 МВт	Собственный интерконнект. Водяное охлаждение
06/2013 , Tianhe-2, КНР, UDT, Inspur [11] ▷ ≈3.1М ядер , 32K Intel Ivy Bridge + 48K Intel Xeon Phi MIC	33.9 Pflops, 24 МВт	Собственный интерконнект. Водяное охлаждение
11/2016 , Sunway TaihuLight, КНР, National Supercomputing Center in Wuxi, NRCPC, Inspur [12] ▷ ≈10.65М ядер , CPU Sunway SW26010 260C 1.45 GHz, 40,960×(4+256 ядер)	93.0 Pflops, 15.4 МВт	Собственный интерконнект. Водяное охлаждение
06/2019 , Summit, США, IBM/NVIDIA [13, 14] ▷ ≈2.4М ядер , IBM POWER9 22C 3.07 GHz – ≈203K ядер, NVIDIA Volta GV100 – ≈2.2М ядер	148 Pflops, 9.8 МВт	Интерконнект: Dual-Rail EDR Infiniband. Водяное охлаждение
6/2020 , Fugaku, Япония, Fujitsu, RIKEN [15, ?russian-] ▷ ≈7.63М ядер , 158,976 CPUs Fujitsu ARM A64FX 48C 2.2GHz 48 ядер	442 Pflops, 29.9 МВт	Собственный интерконнект. Водяное охлаждение
6/2022 , Frontier [™] (OLCF-5, HPE Cray EX235a), США, DoE, SC/ORNL, HPE [?russian-bib:13] ▷ ≈8.7М ядер , 9,248 CPUs AMD Optimized 3rd Generation EPYC 64C 2GHz 64 ядра, 36,992 GPUs AMD Instinct MI250X 220 ядер	1,102 Pflops, 21.1 МВт	Собственный интерконнект Slingshot-11. Водяное охлаждение

Анализ таблицы 1 позволяет вскрыть очевидные тенденции и сделать следующие достоверные (на ближайшие годы) предположения об общем облике систем эксафлопсного уровня:

- *число узлов* – 10–50 тысяч, *число ядер* – 5–10 миллионов;
- *электропотребление* – 15–30 МВт;
- *интерконнект* – скорее всего (7/8 = 87% случаев) класса Custom (Proprietary Network, специфичные решения или недоступные как

отдельное коммерческое решение) либо топовый (труднодоступный) стандартный интерконнект, но с доработками производителя ($1/8 = 13\%$ случаев);

- *охлаждение* – или закрытое водяное, в будущем – более перспективное иммерсионное или охлаждение кипением;
- *вычислитель* – в зависимости от назначения – или однородный (50%) или гибридный (50% случаев);
- *процессоры и ускорители* – либо собственные (44%) или топовые (труднодоступные) стандартные (56% случаев).

4. Тенденции развития гибридных архитектур

Гибридными суперкомпьютерами называют системы, в которых вычислительные узлы в дополнение к основным стандартным процессорам оснащены специализированными процессорами – ускорителями. Идея ускорителей в суперкомпьютерной отрасли использовалась давно, но широкое применение гибридные архитектуры нашли только в последнее десятилетие. Поля, детализирующие использование ускорителей в суперкомпьютерах (Accelerator и Accelerator Cores), появились в записях списка Top500 начиная с июня 2011 года. С этого времени мы и будем анализировать гибридные архитектуры, а каждый суперкомпьютер отнесем к следующим классам по производителям используемых в качестве вычислителей ускорителей:

- NONE: не используются ускорители – однородная (не гибридная) архитектура;
- IBM: спецвычислитель IBM PowerXCell 8i;
- AMD: спецвычислители от компании AMD;
- NVIDIA: спецвычислители от компании NVIDIA;
- Intel: спецвычислители от компании Intel – семейства Intel MIC / Intel Xeon Phi;
- MIX: спецвычислители как от компании NVIDIA, так и от компании Intel;
- PEZY: спецвычислители от компании PEZY Computing;
- Other: используются спецвычислители, не упомянутые выше. Сегодня в этом классе только ускорители, разработанные в Китае: Matrix-2000 и Deep Computing Processor.

Рассмотрим 24 редакции рейтинга Top500 с июня 2011 года по ноябрь 2022 года. На рисунке 4 показаны доли различных классов гибридных суперкомпьютеров в рейтинге Top500 для этих 24 редакций. Подписи

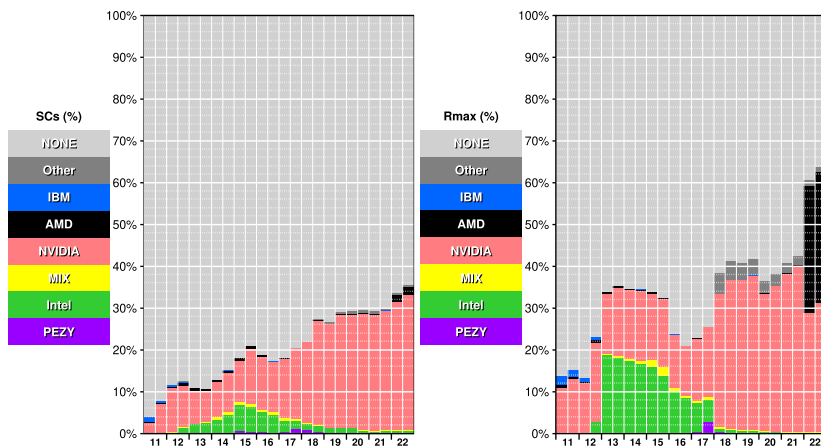


Рисунок 4. Доли различных классов гибридных суперкомпьютеров в рейтинге Top500, редакции с июня 2011 года по июнь 2022 года

по оси абсцисс — года, две столбчатые диаграммы в год соответствуют двум редакциям Top500 в соответствующем году. В левой части рисунка показаны доли в количестве систем (SCs) и за 100% принято 500 систем, а каждая система это 0.2%. В правой части рисунка показаны доли производительности (RMax) и за 100% принята суммарная производительность всего списка Top500 — ΣR_{max} . Можно заметить следующие тенденции:

- по количеству систем в последние годы (2014–2022) гибридные суперкомпьютеры выросли 12.8% ↗ 35.6%, по вкладу в суммарную производительность ΣR_{max} всего списка Top500 (а это и есть истинная доля в отрасли) их доля за этот период выросла 34.4% ↗ 63.8%;
- ускорители компании IBM практически исчезли к 2013 году;
- на пути исчезновения ускорители Intel: доля упала 18.8% ↘ 0.1% ΣR_{max} в период 2013–2022;
- доля NVIDIA выросла 16.4% ↗ 30.9% ΣR_{max} в период 2014–2022;

- ускорители компании AMD практически исчезли к 2017 году, но потом выросли 0.1 ↗ 31.4% $\Sigma R_{\max}$;
- появились интересные новые решения, которые заслуживают внимания (5.4% $\Sigma R_{\max}$ в лучший, 2018, год) – короткие и успешные проекты создания полностью собственных ускорителей, что должно послужить вдохновляющим примером для России:
 - PEZY – 2015, PEZY Computing, Япония;
 - Matrix-2000 – 2017, NUDT, Китай;
 - Deep Computing Processor – 2017, Sugon, Китай.

4.1. Квинтили суперкомпьютеров в списке Top500

Очень важно понимать применимость и пригодность каждой технологии в супер-ЭВМ различной производительности. Например, важно понимать, в каких системах применима и востребована сегодня та или иная технология интерконнекта, тот или иной класс ускорителей и т. д. Для этого поступим следующим образом: разобьем весь список Top500 (а он отсортирован по производительности суперкомпьютеров) на пять групп: квинтиль А – несколько самых мощных суперкомпьютеров из списка; за нею квинтили В, С, D, Е. Размер квинтилей выберем так, чтобы сумма производительности всех суперкомпьютеров, входящих в один квинтиль, была бы близка (как только возможно) к 20% от суммарной производительности $\Sigma R_{\max}$ всего списка Top500. То есть, квинтили А, В, С, D и Е имеют (примерно) равные производительности, но, конечно, содержат разное количество суперкомпьютеров. Для каждого выпуска рейтинга Top500 состав квинтилей А, В, С, D и Е (количество суперкомпьютеров в них) вычисляются заново, исходя из требования: суммарную производительность каждого квинтиля – как можно точнее приблизить к 20% $\Sigma R_{\max}$.

Для редакции Top500 в ноябре 2022 года деление на квинтили получилось таким:

- А: 1 система, Top1, 22,7% $\Sigma R_{\max}$;
- В: 3 системы, Top2–4, 19,0% $\Sigma R_{\max}$;
- С: 22 системы, Top5–26, 19,9% $\Sigma R_{\max}$;
- D: 111 систем, Top27–137, 20,0% $\Sigma R_{\max}$;
- Е: 363 системы, Top138–500, 18,4% $\Sigma R_{\max}$.

4.2. Распределение разных классов гибридных архитектур по квинтилям

Для каждого квинтиля – А, В, С, D, Е – посмотрим доли в нем классов None, Other, IBM, AMD, NVIDIA, MIX, Intel, PEZY (рисунок 5).

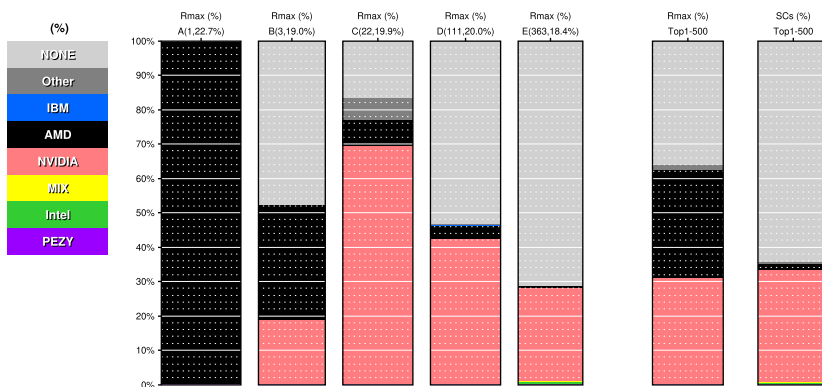


Рисунок 5. Доли различных классов гибридных суперкомпьютеров в суммарной производительности квинтилей А, В, С, D и Е рейтинга Top500 редакции за ноябрь 2022 года

Обращая внимание на форму и центр тяжести у фигур с соответствующим цветом заливки, мы можем сделать следующие суждения для тенденций ноября 2022 года:

- ускорители от AMD используются преимущественно в самых мощных системах – весь квинтиль А и третья часть В;
- ускорители от NVIDIA сосредоточены в большей степени в квинтилях С и D;
- негибридные системы (класс NONE) занимают существенную часть квинтилей В, D, Е;
- класс Other (китайские ускорители) представлены в квинтиле С (6%);
- ускорители классов Intel и MIX совсем слабо представлены в квинтиле Е – покидают рейтинг и отрасль;
- ускорители класса IBM и PEZY покинули отрасль.

5. Тенденции в технологиях интерконнекта

Интерконнект связывает все вычислительные узлы суперкомпьютера в единую систему. Сетевые технологии, используемые в интерконнекте, чаще всего отличаются от сетевых технологий, используемых в локальных (региональных и глобальных) компьютерных сетях, в силу специфических требований, предъявляемых к интерконнекту: не только высокая пропускная способность, но и максимально низкая задержка и максимально высокий темп выдачи сообщений, и еще ряд дополнительных, специфических требований.

В суперкомпьютерной отрасли это определяет серьезную роль технологий интерконнекта. В перечне [16] 10 критических технологических вызовов, стоящих перед отраслью на пути к эксафлопсным системам, разработка перспективных технологий интерконнекта занимает весьма высокое (второе) место.

Поля, детализирующие использование различных технологий интерконнекта в суперкомпьютерах, присутствуют во всех редакциях списка Top500. Опираясь на данные в этих полях, каждый суперкомпьютер во всех редакциях списка Top500 отнесем к различным классам в зависимости от семейства решений, используемых для реализации интерконнекта:

- Infiniband, если используются решения семейства Infiniband. Такие суперкомпьютеры встречаются в Top500 с июня 2003 года до сего дня, а технология реализована в нескольких версиях, с различными техническими характеристиками: Infiniband SDR (8 Гбит/с), Infiniband DDR (16 Гбит/с), Infiniband QDR (32 Гбит/с), Infiniband FDR (54 Гбит/с), Infiniband EDR (100 Гбит/с), Infiniband HDR (200 Гбит/с).
- Ethernet, если используются решения семейства Ethernet. Такие суперкомпьютеры встречаются в Top500 с июня 1996 года до сего дня, а технология реализована в нескольких версиях, с различными техническими характеристиками – от FastEthernet (100 мбит/с) до 100G Ethernet (100 Гбит/с).
- Myrinet, если используются решения семейства Myrinet. Такие суперкомпьютеры встречаются в Top500 с ноября 1998 года до ноября 2020 года.
- Quadrics, если используются решения семейства Quadrics. Такие суперкомпьютеры встречаются в Top500 с июня 1999 года до ноября 2011 года.

- SCI, если используются решения семейства SCI. Такие суперкомпьютеры встречаются в Top500 с июня 2002 года до ноября 2004 года.
- OpTsIntel, если используются решения от компании Intel семейств Omni-Path / TrueScale. Такие суперкомпьютеры встречаются в Top500 с ноября 2013 года до сего дня.
- Custom, если используются решения специфичные (Proprietary Network) или недоступные как отдельное коммерческое решение – доступное только в составе законченной системы.

Подчеркнем отличие между классами:

- Infiniband, Ethernet, Myrinet, Quadrics, SCI – эти технологии интерконнекта доступны как отдельные коммерческие продукты. Любой разработчик своего собственного суперкомпьютера (например, со своей архитектурой вычислительных узлов, с теми или иными процессорами) может купить и применить любое из этих решений.
- Custom – эти технологии интерконнекта не доступны как отдельные коммерческие продукты. Либо недоступны вообще, либо доступны только в составе законченной системы от единственного поставщика.
- OpTsIntel – технологии Intel семейств Omni-Path / TrueScale занимают промежуточную позицию. С одной стороны, эти технологии интерконнекта доступны как коммерческие продукты. С другой стороны, эти решения невозможно применить в суперкомпьютере, если вычислительный узел не имеет соответствующие архитектуру и процессор от компании Intel. Эта ситуация очень близка к ситуации «интерконнект доступен только в составе законченной системы от единственного поставщика» – близка по жесткости технологической зависимости от единственного разработчика и поставщика, зависимости от компании Intel.

Технология компании Intel семейств Omni-Path / TrueScale имела в момент появления весьма серьезные технические характеристики, и это ей позволило успешно войти в отрасль и отвоевать в короткие сроки долю в 11.0% $\Sigma R_{\max}$ в ноябре 2019 года. Слабой стороной решения OpTsIntel является жесткая технологическая зависимость всей системы от компании Intel. И сегодня эта технология покидает отрасль.

Рисунок 6 показывает доли производительности суперкомпьютеров с различными интерконнектами в суммарной производительности $\Sigma R_{\max}$ рейтинга Top500 для 60 редакций с июня 1993 года по ноябрь 2022

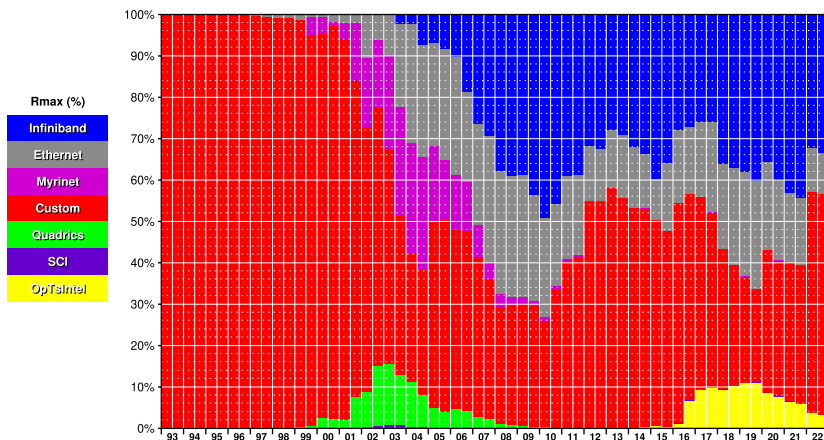


Рисунок 6. Доли суперкомпьютеров с различными интерконнектами в суммарной производительности рейтинга Top500, с июня 1993 по ноябрь 2022 года

года. Подписи по оси абсцисс – года, две столбчатые диаграммы в год соответствуют двум редакциям этого года.

Анализируя рисунок, можно отметить следующие тенденции последних 10 лет:

- технологии, соответствующие классам Myrinet, Quadrics, SCI в последние годы в отрасли не используются. На рисунке хорошо видны годы появления этих решений, годы «рассвета» этих технологий, пик популярности и годы «заката». Сегодня актуально обсуждать только 4 класса: Infiniband, Ethernet, OpTsIntel и Custom.
- с 2012 года самая большая доля суммарной производительности Top500 у суперкомпьютеров с интерконнектом класса Custom. Динамика, 2012–2017–2022:
 - Custom 54.7% ↘ 42.4% ↗ 53.4%.
- следующей по величине (по значимости) является доля Infiniband. Динамика, 2012–2017–2022:
 - Infiniband 32.5% ↘ 26.1% ↗ 33.6%.
- затем следует Ethernet. Динамика, 2012–2017–2022:
 - Ethernet 13.2% ↗ 21.9% ↘ 10.0%
- технологии OpTsIntel покидают Top500 и отрасль. Динамика, 2012–2017–2022:

– OpTsIntel 0.0% ↗ 9.6% ↘ 3.2%.

5.1. Распределение разных классов интерконнекта по квинтилям

Очень важно понимать области применимости каждого класса интерконнектов. Для этого рассмотрим распределение долей разных классов интерконнекта по квинтилям. Определение квинтилей дано выше в разделе 4.1.

Для каждого квинтиля – А, В, С, D, Е – посмотрим доли классов Infiniband, Ethernet, Myrinet, Custom, Quadrics, SCI, OpTsIntel – рисунок 7.

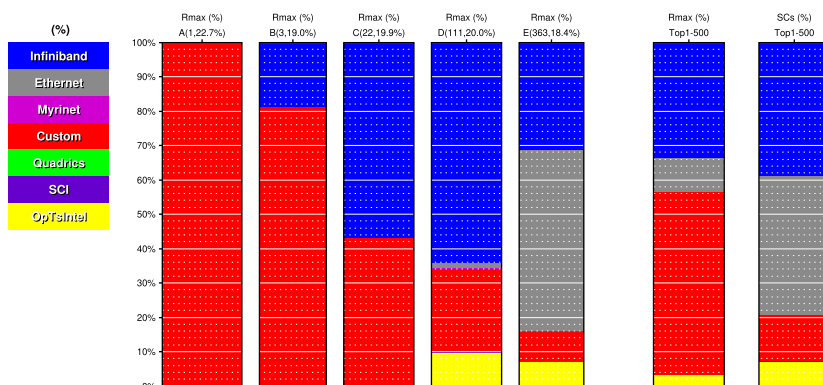


Рисунок 7. Доли суперкомпьютеров с различными интерконнектами в суммарной производительности квинтилей А, В, С, D и Е рейтинга Top500 редакции за ноябрь 2022 года

Видно, что:

- мощнейшие и мощные суперкомпьютеры (квинтили А и В) преимущественно используют интерконнект Custom. Красный цвет заполняет треугольник, который расширяется в направлении самых мощных систем;
- технологию Infiniband преимущественно используют для систем среднего и нижних уровней – самые большие доли в квинтилях С и D. Синий треугольник имеет центр тяжести в этой зоне;
- технологию Ethernet чаще используют в самых слабых системах (D, Е). Серый треугольник существенно расширяется в сторону самых слабых систем;
- решения OpTsIntel покидают рейтинг и отрасль.

Несомненным выводом из рисунка 7 является следующее: разработка российских суперкомпьютеров высшей производительности, скорее всего, не сможет обойтись без интерконнекта класса Custom. А так как его невозможно купить, то, с неизбежностью, его придется разработать самим. Это уже не раз в России весьма успешно делалось: ИПС имени А. К. Айламазяна РАН (российские интерконнекты SKIF 3D-torus и Паутина), НИЦЭВТ (Ангара), ВНИИЭФ (СМПО), НИИ «Квант» и ИПМ им. М. В. Келдыша РАН (МВС-Экспресс).

Отметим, что единственный отечественный интерконнект, который был не только разработан, но и был использован в системе, вошедшей в рейтинг Top500 – это SKIF 3D-torus (ИПС имени А. К. Айламазяна РАН).

6. Тенденции в технологиях центральных процессоров

Поля, детализирующие использование различных технологий центральных процессоров в суперкомпьютерах, присутствуют во всех редакциях списка Top500. Опираясь на данные в этих полях, каждый суперкомпьютер во всех редакциях списка Top500 отнесем к следующим классам, в зависимости от компании-производителя или от используемых в CPU технологий: Alpha, AMD, ARM, CRAY, Fujitsu, Hitachi, HP, IBM, Intel, MIPS, NEC, SPARC. Если процессор не может быть отнесен к одному из данных классов, то он классифицируется как Proprietary – такое имеет место быть, например, по отношению к современным собственным процессорам Китая.

Рисунок 8 показывает доли производительности суперкомпьютеров с различными центральными процессорами в суммарной производительности $\Sigma R_{\max}$ рейтинга Top500 для 60 редакций с июня 1993 года по ноябрь 2022 года. Подписи по оси абсцисс – года, две столбчатые диаграммы в год соответствуют двум редакциям этого года.

Анализируя рисунок 8, можно отметить следующие тенденции:

- с 2020 года в отрасли, по сути, используются процессоры только из пяти классов: IBM, ARM (система Fugaku, Fujitsu), Proprietary (Китай), AMD и Intel;
- решения IBM за период июнь 2020–ноябрь 2022 сильно потеряли свои позиции:
 - IBM 28.7% $\searrow$ 6.4%.
- система Fugaku (компании Fujitsu) показала, что однородные системы на ARM-процессорах могут иметь рекордные мировые показатели и быть эффективными для широкого спектра задач, включая реализацию искусственных нейронных сетей. Отметим, что наиболее близкими разработками в России являются процессоры семейства Байкал. Динамика доли за период июнь 2020–ноябрь 2022:

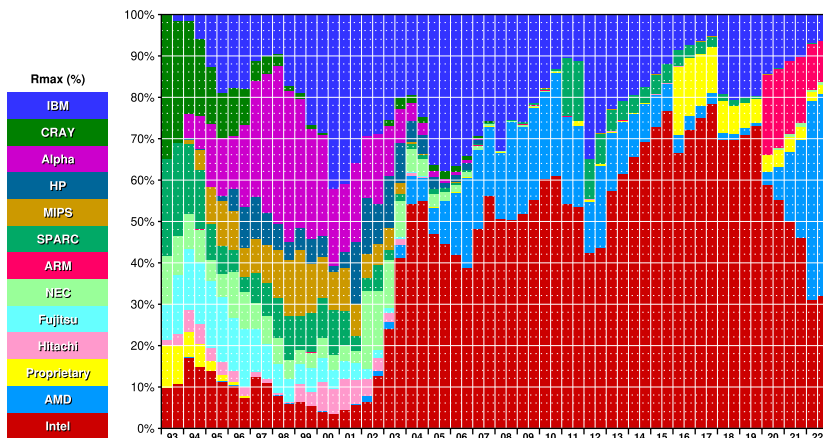


Рисунок 8. Доли суперкомпьютеров с различными центральными процессорами в суммарной производительности рейтинга Top500, редакции с июня 1993 года по ноябрь 2022 года

– ARM 19.2% ↘ 10.0%

- китайские процессоры с 2016 г. совпадают с классом Proprietary. Отметим: по всей видимости, Китай взял курс на сокрытие достаточно большого числа своих крупных установок. В публичной сфере динамика доли китайских процессоров в период июнь 2016–ноябрь 2022:

– Proprietary 16.5% ↘ 2.0%

- ожидаемо самые крупные доли у двух классов центральных процессоров: AMD и Intel. И за период ноябрь 2019–ноябрь 2022 AMD радикально расширил свою долю, отняв ее, по сути, у Intel. Данное событие, как представляется, связано с давней стратегией развития AMD как fabless-компании (в контрасте с Intel) и в нескольких удачных архитектурных разработках AMD. Динамика за ноябрь 2019–ноябрь 2022:

– AMD 0.8% ↗ 48.8%

– Intel 72.9% ↘ 32.0%

6.1. Распределение разных классов центральных процессоров по квинтилям

Рассмотрим распределение долей разных классов центральных процессоров по квинтилям. Определение квинтелей дано в разделе 4.1.

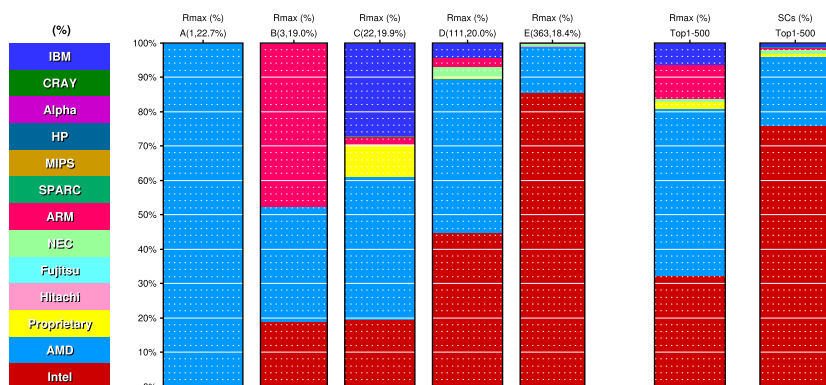


Рисунок 9. Доли суперкомпьютеров с различными центральными процессорами в суммарной производительности квинтилей А, В, С, D и Е рейтинга Top500 редакции за ноябрь 2022 года

Рисунок 9 показывает доли центральных процессоров для каждого квинтиля – А, В, С, D, Е. Видно, что:

- классы центральных процессоров AMD и ARM по большей части представлены в мощнейших и мощных суперкомпьютерах (квинтили А и В), класс IBM представлен в квинтилях С и D с явным тяготением к старшим системам;
- класс Intel представлен в квинтилях (В, С, D, Е), но чаще всего используется в слабых системах – бордовый треугольник существенно расширяется в сторону самых слабых систем;
- остальные классы (Proprietary, NEC) не интересны для серьезного анализа.

Вывод: для построения систем высшей производительности перспективными центральными процессорами являются либо топовые (значит ограничено доступные) модели самых успешных коммерческих разработок, – сегодня это процессоры AMD, – либо собственные процессоры, на базе одной из доступных архитектур, таких как ARM (доказал свою перспективность для использования в суперкомпьютерах) и, возможно, RISC-V.

7. Компании-производители суперкомпьютеров

Поле «Manufacturer» присутствует во всех доступных редакциях списка Top500. Опираясь на данные в этом поле, каждый суперкомпьютер

во всех редакциях списка Top500 отнесем к одному из следующих классов, в зависимости от компании-производителя суперкомпьютера: Cray, IBM, Lenovo, Inspur, NRCPC, NUDT, Sugon, Fujitsu, NVIDIA, Atos, HP, Other. Классы выбраны и отсортированы так, чтобы:

- выявлялись все основные участники отрасли, все второстепенные отнесены к классу Other;
- максимально учитывались изменения последних годов в отрасли – например, агрессивное вторжение в отрасль китайских компаний-производителей (Lenovo, Inspur, NRCPC, NUDT, Sugon), на диаграммах эти классы расположены рядом и для них используются цвета «бледной» палитры;
- при переименовании компании на протяжении 1993–2022 гг., класс сохраняет свое имя. Например, Hewlett-Packard до недавнего времени обозначал себя в данных Top500 как HP или Hewlett-Packard, а теперь HPE (Hewlett-Packard Enterprise) – за классом сохранили имя HP.

Рисунок 10 показывает доли производительности суперкомпьютеров от различных производителей в суммарной производительности ΣR_{max} рейтинга Top500 для 60 редакций с июня 1993 года по ноябрь 2022 года. Подписи по оси абсцисс – года, две столбчатые диаграммы в год соответствуют двум редакциям этого года.

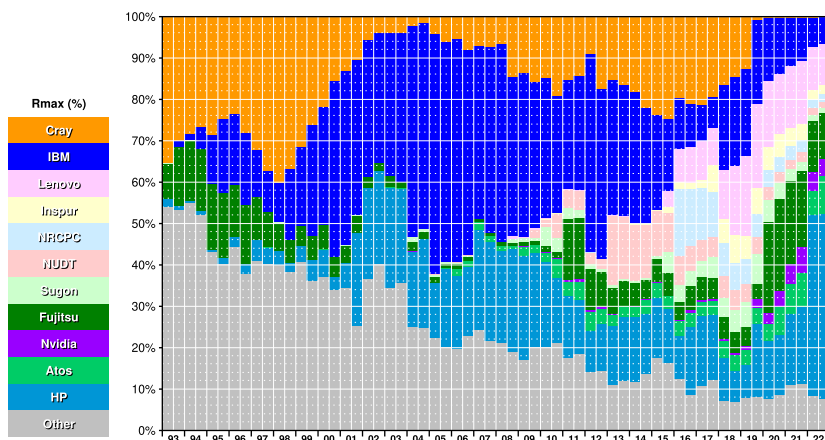


Рисунок 10. Доли суперкомпьютеров различных производителей в суммарной производительности рейтинга Top500, с июня 1993 года по ноябрь 2022 года

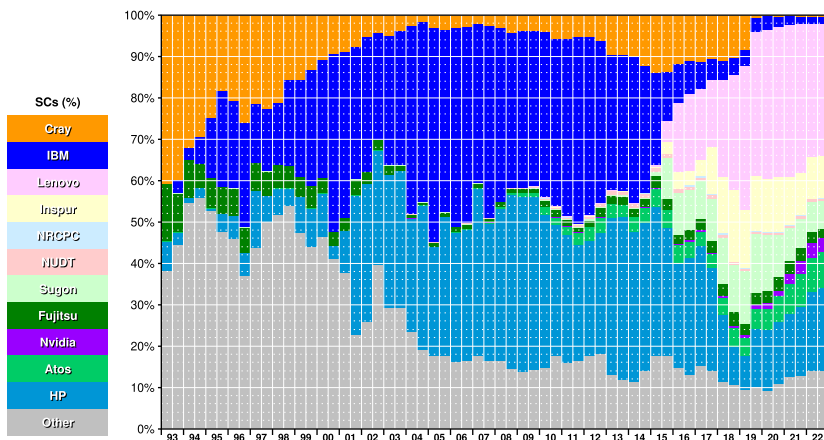


Рисунок 11. Доли количества суперкомпьютеров от различных производителей в общем количестве (500) рейтинга Top500, с июня 1993 года по ноябрь 2022 года

Рисунок 11 показывает доли систем (одна система – это 0.2% от 500) суперкомпьютеров от различных производителей в Top500 для 60 редакций с июня 1993 года по ноябрь 2022 года.

Тем самым, рисунок 10 иллюстрирует реальные объемы продаж, а рисунок 11 – количество продаж. Разница между ними очевидна: много мелких сделок в сумме не дадут большой объем.

Анализируя эти рисунки, можно отметить следующие самые яркие события за последние годы:

- агрессивное вторжение китайских производителей в отрасли с огромным количеством сделок по всему миру и достаточно большим объемом этих сделок. При этом необходимо помнить, что Китай в последние годы закрывает сведения о своих крупных системах. Показатели за июнь 2019–ноябрь 2022:
 - доля SCs 62.2% ↘ 49.6%;
 - доля ΣR_{max} 41.3% ↘ 16.6%;
- Hewlett-Packard купил Cray. Как результат, Hewlett-Packard приобрел собственный интерконнект с рекордными характеристиками и в классе HP появились самые крупные системы (чего раньше никогда не было). Доли числа сделок изменились не существенно, а объемы сделок – радикально. Показатели за июнь 2019–ноябрь 2022:
 - доля SCs 8.2% ↗ 20.2%;

- доля $\Sigma R_{\max}$ 7.9% ↗ 44.5%;
- Компания NVIDIA купила Mellanox. Как результат, NVIDIA приобрела собственный набор технологий и компетенций для того, чтобы выступать в роли компании-поставщика законченных систем, а не только GPU для них. Показатели за июнь 2019–ноябрь 2022:
 - доля SCs 0.8% ↗ 3.4%;
 - доля $\Sigma R_{\max}$ 1.1% ↗ 4.1%.

7.1. Распределение компаний-производителей суперкомпьютеров по квинтилям

Рассмотрим распределение долей разных компаний-производителей по квинтилям. Определение квинтилей дано в разделе 4.1. Рисунок 12 показывает доли компаний-производителей для каждого квинтиля – А, В, С, D, Е.

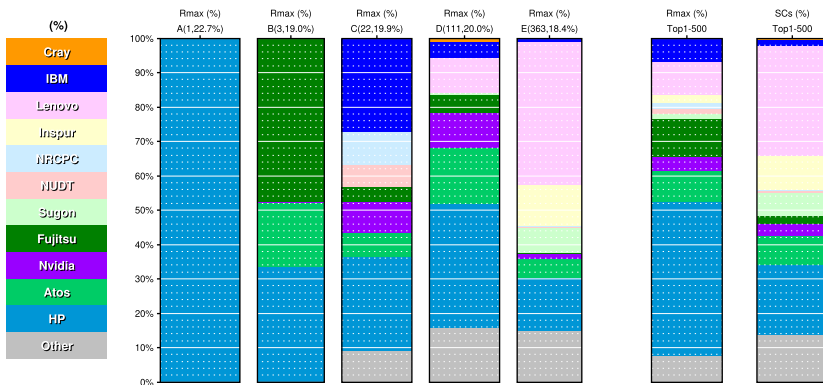


Рисунок 12. Доли суперкомпьютеров различных компаний-производителей в суммарной производительности квинтилей А, В, С, D и Е рейтинга Top500 редакции за ноябрь 2022 года

Видно, что:

- Hewlett-Packard занимает значительные доли не только в квинтилях С, D, Е (что было давно, это традиционная позиция компании), но и в самых высоких сегментах – А, В – что стало возможным после приобретения компании Cray;
- IBM и Fujitsu хорошо представлены в квинтилях среднего сегмента (В, С, D). Их цвета существенно расширяются в сторону крупных систем;

- китайские компании (Lenovo, Inspur, NRCPC, NUDT, Sugon) представлены в начальном сегменте (C, D, E), их цвета существенно расширяются в сторону слабейших систем.

7.2. Основные выводы из анализа компаний-производителей суперкомпьютеров

Среди компаний-производителей суперкомпьютеров произошли существенные изменения в последние годы. Китайские производители осуществили серьезную интервенцию в отрасли: на их долю в ноябре 2019 года приходилось 63.0% количества сделок и 43.4% объемов (доли $\Sigma R_{\max}$). Больше всего доля Китая среди систем начального уровня. Для построения систем высшей производительности важен факт консолидации крупнейших участников отрасли:

- Hewlett-Packard купил Cray. Как результат, Hewlett-Packard приобрел собственный интерконнект с рекордными характеристиками, и в классе HP появились самые крупные системы (чего раньше никогда не было);
- компания NVIDIA купила Mellanox. Как результат, NVIDIA приобрела собственный набор технологий и компетенций для того, чтобы выступать в роли компании-поставщика законченных систем, а не только GPU для них.

Доли числа сделок этих двух компаний изменилась значительно (в 2.6 раза), а объемы сделок – вообще радикально (в 5.4 раза), так как эта консолидация повлияла именно на самые крупные проекты, динамика июнь 2019–ноябрь 2022:

- доля SCs 9.0%; ↗ 23.6%;
- доля $\Sigma R_{\max}$ 9.0%; ↗ 48.6%.

8. Области применения суперкомпьютеров

Поля («Application Area» и «Segment»), позволяющие определить область применения суперкомпьютеров, присутствуют во всех редакциях списка Top500. Опираясь на данные в этих полях, каждый суперкомпьютер во всех редакциях списка Top500 отнесем к одному из следующих классов:

- RnD – суперкомпьютер используется для исследований и разработок; для генерации новых знаний, технологий, материалов; для научных расчетов в НИИ и университетах;
- Industry – суперкомпьютер используется напрямую для задач реальной экономики: инженерные расчёты, финансовые приложения, услуги связи и интернет-сервисы и т. п.;

- Gov.Mil – суперкомпьютер используется в государственных интересах или для военных нужд;
- Unknown – недостаточно информации для отнесения к одному из классов выше.

В качестве примера использования суперкомпьютерной отрасли в государственных (военных) целях, приведем программу США «Crossroads» (Лос-Аламосская Национальная Лаборатория, создана в интересах Программы управления запасами (SSP) и Национального управления по ядерной безопасности (NNSA), министерство энергетики США) [17]. Ее цель: изучение и использование новых технологий для повышения эффективности приложений и рабочих процессов.

Crossroads, посредством симуляций на супер-ЭВМ, создаст способность прогнозирования ключевых физических явлений, количественную оценку и агрегирование неопределенностей в инструментах моделирования, при этом, обеспечивая точность физики, численную достоверность и геометрическую точность.

Рисунок 13 показывает доли производительности суперкомпьютеров с разными областями их применения в суммарной производительности рейтинга $\Sigma R_{\max}$ для 60 редакций с июня 1993 года по ноябрь 2022 года.

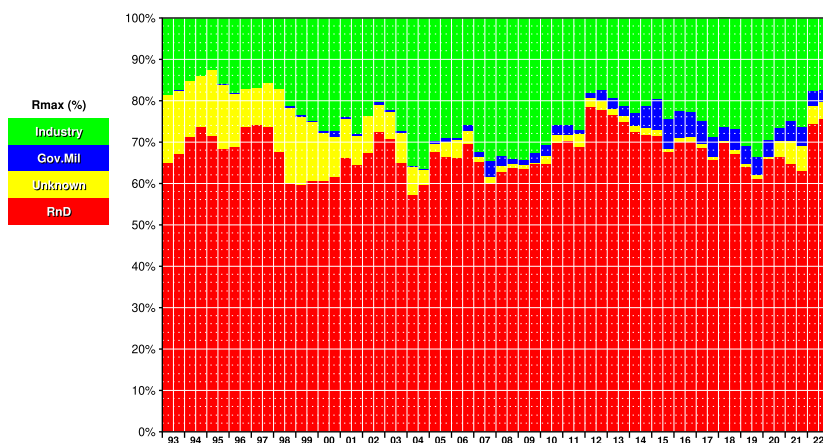


Рисунок 13. Доли суперкомпьютеров с разными областями их применения в суммарной производительности рейтинга Top500, редакции с июня 1993 года по ноябрь 2022 года

Подписи по оси абсцисс – года, две столбчатые диаграммы в год соответствуют двум редакциям списка Top500 этого года.

Рисунок 14 показывает доли систем (доли в SCs – одна система – это 0.2% от 500) суперкомпьютеров с разными областями их применения в рейтинге Top500 для 60 редакций с июня 1993 года по ноябрь 2022 года.

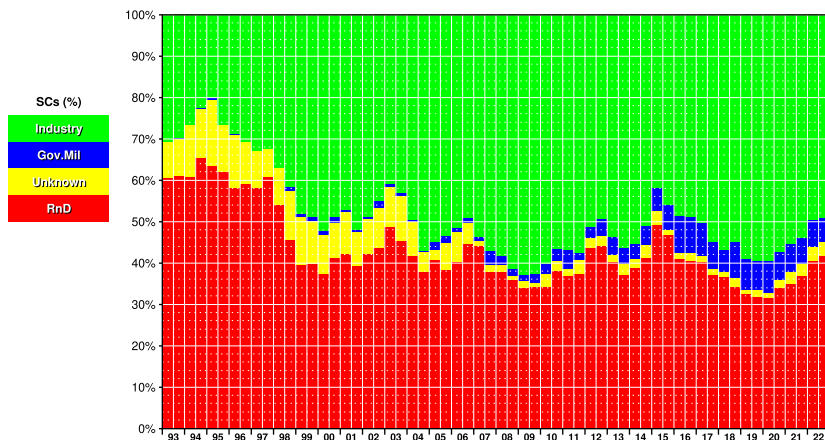


РИСУНОК 14. Доли количества суперкомпьютеров с разными областями их применения в общем количестве (500) рейтинга Top500, редакции с июня 1993 года по ноябрь 2022 года

Анализируя рисунки, можно отметить следующее:

- классы Gov.Mil и Unknown имеют небольшие доли и в SCs, и в ΣR_{max} . Основные классы применения суперкомпьютеров это Industry и RnD;
- во все времена подавляющая доля производительности (доля ΣR_{max}) суперкомпьютеров использовалась для генерации новых знаний (класс RnD), а не для непосредственных прикладных задач реальной экономики (класс Industry). За последние пять лет доля ΣR_{max} Industry 26.0% в среднем, доля ΣR_{max} RnD 67.1% в среднем;
- во все времена подавляющая доля количества (доля SCs) суперкомпьютеров использовалась для непосредственных прикладных задач реальной экономики (класс Industry), а не для генерации новых знаний (класс RnD). За последние пять лет доля SCs Industry 55.5% в среднем, доля SCs RnD 35.4% в среднем;
- таким образом, очевидно, что в классе Industry – как правило, слабые системы, а в классе RnD – крупные системы. Более наглядно это можно увидеть при анализе квинтилей.

8.1. Распределение областей применения суперкомпьютеров по квинтилям

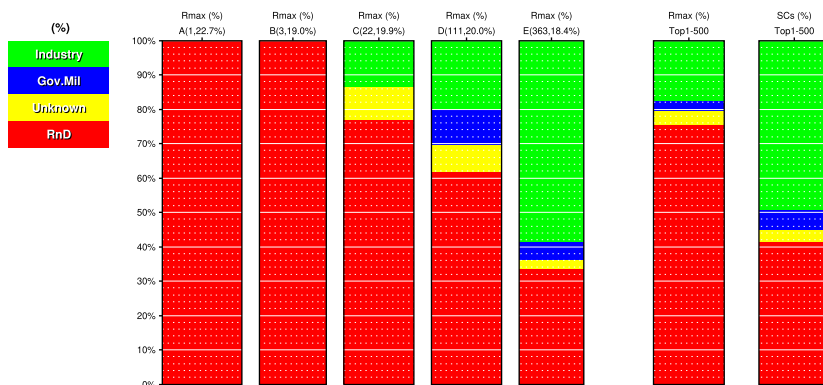


Рисунок 15. Доли областей применения суперкомпьютеров в суммарной производительности квинтилей А, В, С, D и E рейтинга Top500 редакции за ноябрь 2022 года

Рисунок 15 показывает распределение долей разных областей применения суперкомпьютеров по квинтилям А, В, С, D, E в ноябре 2022 года. Видно, что:

- самые крупные системы (квинтили А, В) используются исключительно для генерации новых знаний – класс RnD. Цвет этого класса представлен во всех квинтилях, он явно и существенно расширяется к самым крупным системам;
- цвет класса Industry представлен в квинтилях С, D, E и он явно и существенно расширяется к самым слабым системам. Самые слабые системы (квинтиль E) преимущественно используются для прикладных задач реальной экономики;
- суперкомпьютеры классов Gov.Mil и Other представлены не широко и только в квинтилях С, D, E.

8.2. Основные выводы из анализа областей применения суперкомпьютеров

Для построения систем высшей производительности важен следующий факт – такие системы:

- всегда используются для исследований и разработок; для генерации новых знаний, технологий, материалов; для научных расчётов;
- никогда не используются напрямую для задач реальной экономики: инженерных расчетов, финансовых приложений, услуг связи, интернет-сервисов и т. п.

Это накладывает свой отпечаток на класс решаемых задач: по всей видимости, он достаточно широкий и порой непредсказуемый – на самом переднем крае научного поиска за время жизни рекордной установки (около 5–7 лет) могут возникнуть самые неожиданные задачи, требующие для решения самые неожиданные математические и алгоритмические инструменты. Это следует учитывать при разработке архитектурных решений в части и аппаратных, и программных средств таких систем.

9. Состояние суперкомпьютерной инфраструктуры в мире и в России

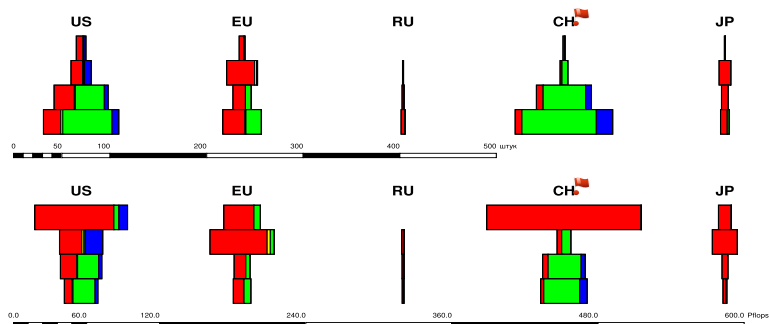
9.1. Суперкомпьютерная инфраструктура стран-лидеров и России

Анализ уровня развития и освоения СКТ в России и в других странах – США, Китай, Япония, Евросоюз – начнем с обзора их суперкомпьютерной инфраструктуры – СК-инфраструктуры.

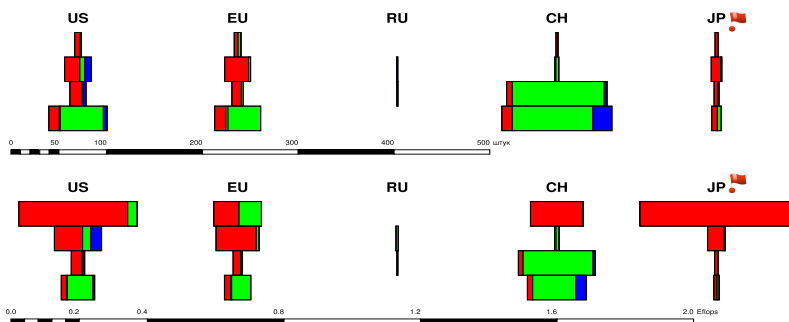
Здесь важно отметить, сколько суперкомпьютеров (мы можем анализировать только публичные суперкомпьютеры) того или иного уровня имеется в стране, какая их суммарная производительность, в каких областях они используются. Будем выделять четыре уровня СК-инфраструктуры:

- суперкомпьютеры уровня Top1–20 (то есть, занимающие первые 20 мест в рейтинге). На практике, эти системы устанавливаются в крупнейших национальных суперкомпьютерных центрах (СКЦ);
- суперкомпьютеры уровня Top21–100 – крупнейшие региональные и отраслевые СКЦ;
- суперкомпьютеры уровня Top101–250, – крупные региональные и корпоративные СКЦ;
- суперкомпьютеры уровня Top251–500, – СКЦ предприятий и научных учреждений.

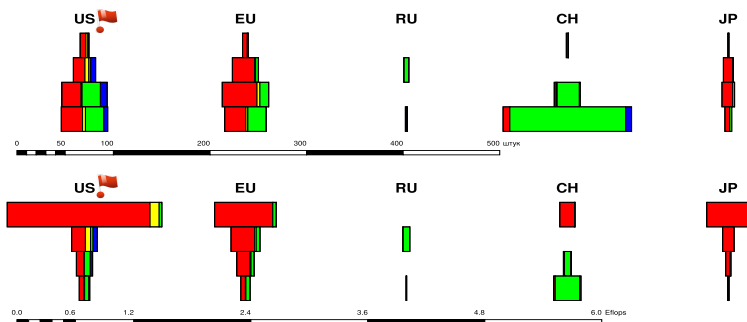
Рисунки 16 показывают характеристики СК-инфраструктур США, Евросоюза, России, Китая и Японии в разные моменты времени.



(а) в июне 2016 года



(б) в июне 2020 года



(в) в ноябре 2022 года

Рисунок 16. Суперкомпьютерная киберинфраструктура США (US), стран Евросоюза (EU), России (RU), Китая (CH) и Японии (JP)

Поясним рисунки 16.

- в верхней части каждого рисунка для каждой страны размер каждого слоя «пирамиды» указывает количество суперкомпьютеров каждого из четырёх уровней, в нижней части показана суммарная производительность суперкомпьютеров этих уровней. Красный флажок отмечает страну расположения системы Top 1;
- цветами выделены доли декларированных областей применения СКЦ: красный цвет – исследования и разработки; зеленый цвет – индустриальные применения в реальной экономике; синий цвет – государственные нужды и оборона; желтый цвет – нет информации.

Очевидно следующее:

- все страны (кроме России) имеют сбалансированную СК-инфраструктуру – в заметном количестве присутствуют СКЦ всех 4 уровней;
- Россия, по сути, не обладает никакой СК-инфраструктурой. По крайней мере, такой, которая отвечала бы ее планам по развитию цифровой экономики.

9.2. Суперкомпьютеры, установленные в Российской Федерации

На ноябрь 2022 года в списке Top500 имелись 7 машин, установленных в России. Одна из них – суперкомпьютер известной НРС-компании Т-Платформы¹:

- 290 место – Lomonosov 2 (T-Platform A-Class Cluster, Xeon E5-2697v3 14C 2.6GHz, Intel Xeon Gold 6126, Infiniband FDR, Nvidia K40m/P-100, T-Platforms, Moscow State University – Research Computing Center), 2.48 PFlops – 0.22% от Top1, слабее в 455 раз, чем Top1.

Эта машина входит в список Top500 с 2014 года и *доступна для использования*^{URL} в исследовательских проектах в области медицины, инженерных расчетов, астрофизики, материаловедения, энергетики и других областей.

Три суперкомпьютера компании Яндекс^{URL}:

- 25 место – Chervonenkis (YANDEX, AMD EPYC 7702 64C 2GHz, NVIDIA A100 80GB, Infiniband), 21.53 PFlops – 1.95% от Top1, слабее в 51.3 раза, чем Top1;

¹Новый суперкомпьютер в МГУ вошел в Top500^{URL} – новость от 19.11.2014 издательства «Открытые системы»

- 44 место – Galushkin (YANDEX, AMD EPYC 7702 64C 2GHz, NVIDIA A100 80GB, Infiniband), 16 PFlops – 1.45% от Top1, слабее в 69 раз, чем Top1;
- 47 место – Lyapunov (Inspur NF5488A5, AMD EPYC 7662 64C 2GHz, NVIDIA A100 40GB, Infiniband, NVIDIA, Inspur), 12.8 PFlops – 1.16% от Top1, слабее в 86.2 раза, чем Top1.

Данные суперкомпьютеры используются для задач машинного обучения в дата-центрах Яндекса, которые запускаются через разработанную в Яндексе платформу YT для хранения и обработки больших данных. К задачам относятся: нейросетевые модели поиска, сервисы Алисы и другие сервисы компании. На фоне представленных машин в Top500, компания Яндекс продвигает свой сервис для ML-разработки – Yandex DataSphere в рамках инфраструктуры Yandex Cloud. Отметим, что производителями машин Chervonenkis и Galushkin являются компании YANDEX, NVIDIA (США), а машина Lyapunov произведена компаниями NVIDIA, Inspur (Китай).

Две машины компании Сбербанк:

- 50 место – Christofari Neo (NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 80GB, Infiniband, SberCloud), 11.95 PFlops – 1.08% от Top1, слабее в 92.6 раза, чем Top1;
- 86 место – Christofari (NVIDIA DGX-2, Xeon Platinum 8168 24C 2.7GHz, Mellanox InfiniBand EDR, NVIDIA Tesla V100, SberCloud), 6.6 PFlops – 0.61% от Top1, слабее в 164 раза, чем Top1.

Сбербанк использует эти машины как для внутренних задач компании (распознавание речи, генерация голоса и др.), так и для предоставления облачного сервиса ML Space, обеспечивающего работу с ML: решаются задачи государственных и коммерческих заказчиков, а также научно-исследовательских групп. В настоящий момент, на [сайте компании](#)^{URL} можно подать заявку на тестовый доступ к облачной платформе в рамках SberCloud. Производителем машин Christofari Neo и Christofari является компания NVIDIA (США).

Компания МТС представила *одну машину в списке*^{URL}:

- 352 место – MTS GROM (NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 40GB, Infiniband, Nvidia, CloudMTS) 2.26 PFlops – 0.2% от Top1, слабее в 500 раз, чем Top1.

Суперкомпьютер планируется использовать в рамках облачной инфраструктуры компании CloudMTS. Ожидается, что на кластере будут считаться как собственные задачи компании, так и задачи для крупного бизнеса, что позволит кратно сократить время на разработку и внедрение проектов на основе искусственного интеллекта и больших данных. Производителем данной машины является компания NVIDIA (США).

Таким образом, государственная инфраструктура в Российской Федерации слабо развита, большая часть существующих машин из Top500, описанных выше, работают в собственных интересах корпораций. Почти все машины, установленные в России и входящие в Top500 (за исключением Lomonosov 2) – либо совместного производства с зарубежными компаниями (США, Китай), либо полностью зарубежные.

9.3. Китай начал скрывать свои суперкомпьютерные разработки

Отдельные слова стоит сказать про Китай. Есть серьезные основания для того, чтобы утверждать, что Китай начал скрывать сведения про свои самые мощные системы. И, по всей видимости, в Китае доля непубличных суперкомпьютеров (и непубличная доля производительности) стали резко превосходить то, что мы имеем для других развитых стран.

Основанием для того, чтобы сделать такой вывод, послужили следующие факты:

- еще летом 2021 года появились сообщения [18] о вероятном существовании в Китае двух суперкомпьютеров с производительностью более 1 EFlops. Официальных подтверждений или опровержений не последовало;
- США внес² в «черный список» ряд суперкомпьютерных организаций Китая. Конечно, это стимулирует Китай скрывать свои успехи, чтобы не давать поводов для расширения санкционного списка;
- если сравнить динамику развития СК-инфраструктур ведущих стран за несколько лет на рисунках 16а, 16б и 16в, будет видно, что у Китая была серьезная СК-инфраструктура в 2017–2020 годах – пожалуй, даже более развитая, чем у США. И это все это было достигнуто Китаем всего за 8 лет – в 2009 году практически никакой

²по сообщению Карасёва С. на [ServerNews[®]](#) от 09.04.2021 «США внесли в чёрный список ряд суперкомпьютерных центров Китая»

СК-инфраструктуры Китай не имел. За восемь лет он догнал и перегнал США. И вдруг, после 2020 года, СК-инфраструктура Китая стала «таять».

9.4. Интегральные показатели степени освоения СКТ странами-лидерами и Россией

Важный интегральный показатель степени освоения СКТ страной является суммарная производительность всех суперкомпьютеров страны. Рисунок 17 иллюстрирует этот показатель и наглядно отображает достижения Китая в последние годы и недопустимое отставание России.

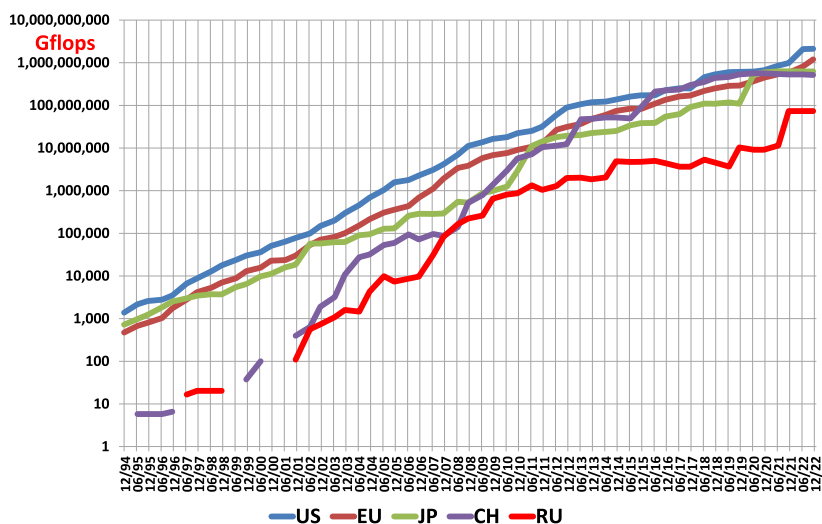


Рисунок 17. Суммарная производительность всех суперкомпьютеров США, стран Евросоюза, России, Китая и Японии за период от июня 1995 года по ноябрь 2022 года

Оценить отставание России (в годах) в области СКТ от других стран (рассмотрим на примере США) можно следующим образом: берем любой момент времени, смотрим суммарную производительность суперкомпьютеров России и дальше смотрим, сколько лет назад такое же (или чуть меньше) значение было у США. Эти вычисления повторяем для каждого момента времени и для всех стран, получая рисунок 18.

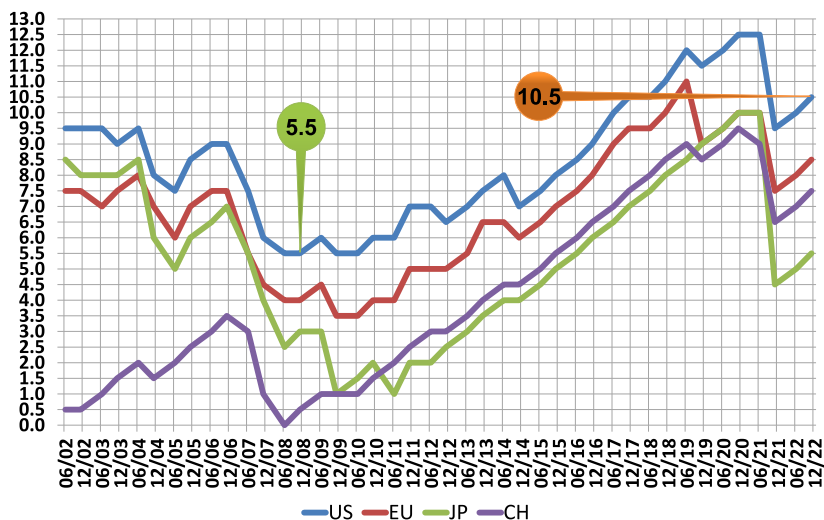


Рисунок 18. Отставание России (в годах) по суммарной производительности суперкомпьютеров от США (US), стран Евросоюза (EU), Китая (CH) и Японии (JP) за период июнь 2002 года по ноябрь 2022 года

Если в 2006–2008 Россия резко сокращала свое отставание от США, Евросоюза, Китая и Японии, то начиная с 2010 года мы практически неуклонно ежегодно наращаем свое отставание от этих стран. И в июне 2021 года это отставание достигло максимума за весь период наблюдения.

Термин «отставание» требует пояснения. Например, сейчас отставание от США – 10.5 лет. Можем ли мы «догнать Америку» за 10.5 лет? Ответ, формально, да можем, например, при реализации одного из следующих сценариев:

- США в своем развитии полностью останавливается, а Россия 10.5 лет подряд развивается с теми же темпами, что до этого были у США. Через 10.5 лет мы сравняемся;
- США продолжает свое развитие, но Россия 10.5 лет подряд развивается с темпами, вдвое большими, чем они были у США за предыдущие годы. Через 10.5 лет мы сравняемся.

Оба сценария маловероятны. Поэтому, наверное, более правильно использовать термин не «отставание», а «задержка в развитии» – недостаток,

который очень тяжело преодолеть.

Аналогичным образом можно объективно и достоверно рассчитать отставание России (в годах) от уровней технологий. рисунок 19 показывает отставание России от уровня технологий Top1, Top5 и Top10.

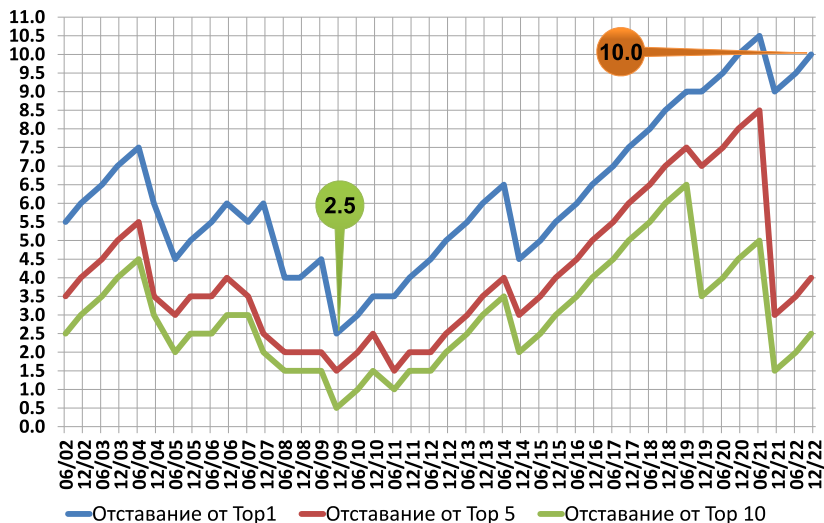


РИСУНОК 19. Отставание (в годах) России от уровней технологий классов Top1, Top5 и Top10 за период июнь 2002 года по ноябрь 2022 года

Рассчитывалось это так: в каждый момент времени рассматриваем самый мощный суперкомпьютер в России и оцениваем, сколько лет назад этот суперкомпьютер занял бы первое (соответственно пятое и десятое) место в рейтинге. Эта величина (в годах) и будет отставанием России от технологий класса Top1 (Top5, Top10).

Видны периоды, когда Россия сокращала технологическое отставание. Было даже время, когда от технологий уровня Top10 нас отделяли всего полгода. Однако в последние годы технологическое отставание неуклонно росло и в июне 2021 года оно достигло максимума за весь период наблюдения.

Наконец, роль страны в мировой суперкомпьютерной отрасли, степень устремленности к цифровой экономике хорошо характеризует доля страны в совокупной мировой вычислительной производительности. То есть,

отношение суммарной производительности $\Sigma R_{\max}$ суперкомпьютеров, установленных в данной стране, к суммарной производительности всего списка Top500 $\Sigma R_{\max}$. По своей роли и своему смыслу этот параметр чем-то аналогичен доле экономики страны в мировом ВВП.

Рисунок 20 показывает долю России в совокупной мировой вычислительной производительности с июня 1999 года по ноябрь 2022 года.

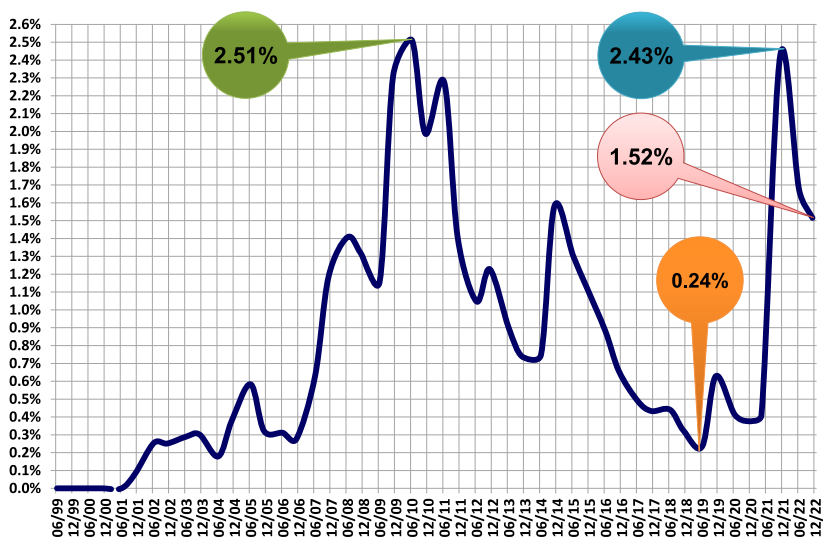


РИСУНОК 20. Доля России в совокупной мировой вычислительной производительности суперкомпьютеров в период с июня 1999 года по ноябрь 2022 года

В лучшие времена (2009 год) эта доля достигала 2.51%, а в последние годы неуклонно сокращалась до 0.24% (июнь 2019 года), но потом за счет закупок шести систем компаниями Яндекс, Сбер и МТС этот показатель немного вырос и сейчас составляет 1.52%.

9.5. Индекс цифровизации стран-лидеров и России

В Таблице 2 вводится ещё один важный показатель. В таблице приведены доли различных стран (на ноябрь 2022 года) в совокупной мировой производительности суперкомпьютеров $\Sigma R_{\max}$ и их доли

в мировой экономике – доли в мировом ВВП, данные о ВВП различных стран взяты из [19], а также отношение этих долей, обозначенное как $K_{ц}$ – индекс цифровизации страны. Можно обосновать (в данной работе опущено), что индекс цифровизации страны оценивает величину «средний объём вычислений, затрачиваемый на выпуск единицы продукции в данной стране».

ТАБЛИЦА 2. Индекс цифровизации стран-лидеров и России.
Ноябрь 2022

Колонки таблицы:				
(А) Доли стран в суммарной производительности супер-ЭВМ мира ΣR_{max} ;				
(В) Доли стран в мировом ВВП;				
(С) Индекс цифровизации, $K_{ц}$ – отношение (А/В);				
(D) Во сколько раз индекс цифровизации данной страны выше индекса цифровизации России.				
Страна	Доля страны в ΣR_{max}	Доля страны в сумме ВВП	Индекс цифровизации, $K_{ц}$ страны	Отношение $K_{ц}$ страны к $K_{ц}$ России
	(А)	(В)	(C = A/B)	(D)
Китай	12.04%	18.45%	—	—
США	47.35%	23.93%	1.98	2.19
Япония	14.23%	5.14%	2.77	3.06
Евросоюз	18.54%	17.78%	1.04	1.15
Россия	1.67%	1.85%	0.91	—
Мир	100.00%	100.00%	1.00	1.10
Четыре лидера	92.17%	65.30%	1.41	1.56

Как легко заметить, у России этот показатель сильно уступает развитым странам. Индекс цифровизации у России хуже в 2.19 раза – чем у США; в 1.15 раз – чем у Евросоюза; в 3.06 раз – чем у Японии; в 1.56 раз – чем у четырех стран-лидеров в совокупности; в 1.1 раза – чем у мира в среднем.

Заключение

Просуммируем выводы всех разделов данной статьи.

Мировая суперкомпьютерная отрасль на рубеже 2008 года, несомненно, столкнулась с научно-техническими сложностями на пути развития СКТ. Начиная с 2008 года изменилась скорость роста достигнутой максимальной производительности: вместо прироста «в 1000 раз за 11 лет» мы имеем прирост «в 1000 раз за 13–17 лет». Это привело к пересмотру будущих прогнозов достижения новых рубежей производительности: 1 Eflops (10^{18} flops) был достигнут в июне 2022 года, а 1 Zflops (10^{21} flops) стоит ожидать в 2035–2039 году.

Анализ самых мощных суперкомпьютеров последнего десятилетия позволяет сделать следующие общие достоверные (на ближайшие годы) предположения об общем облике систем эксафлопсного уровня:

- число узлов – 10–50 тысяч, число ядер – 5–10 миллионов;
- электропотребление – 15–30 МВт;
- интерконнект – скорее всего ($7/8 = 87\%$ случаев) класса Custom (Proprietary Network, специфичные решения или недоступные как отдельное коммерческое решение), либо топовый (труднодоступный) стандартный интерконнект, но с доработками производителя ($1/8 = 13\%$ случаев);
- охлаждение – или закрытое водяное, или более перспективное – иммерсионное, охлаждение кипением и др.;
- вычислитель – в зависимости от назначения – или однородный (50%), или гибридный (50% случаев);
- процессоры и ускорители – либо собственные (44%), либо топовые (труднодоступные) стандартные (56% случаев).

Анализ гибридных архитектур в редакциях Top500 позволяет установить, что вклад гибридных архитектур в общую производительность Top500 вырос и достиг 63.8% ΣR_{max} . Среди актуальных решений для гибридных систем можно отметить ускорители AMD (31.4% ΣR_{max}) и NVIDIA (30.9% ΣR_{max}). Заслуживают внимания интересные новые решения (5.4% ΣR_{max} в лучший, 2018, год) – короткие и успешные проекты создания полностью собственных ускорителей, что должно послужить вдохновляющим примером для России: PEZY (2015, PEZY Computing, Япония), DeepComputingProcessor (2017, Sugon, Китай), Matrix-2000 (2017, NUDT, Китай).

В мире продолжают интенсивно развиваться различные технологии интерконнекта. Они играют исключительно высокую роль для всего

комплекса СКТ. На сегодняшний день, из коммерчески доступных технологий используются решения класса Custom (53.4% $\Sigma R_{\max}$), Infiniband (33.6% $\Sigma R_{\max}$), Ethernet (10.0% $\Sigma R_{\max}$). С 2012 года самая большая доля суммарной производительности Top500 у суперкомпьютеров с интерконнектами, которые недоступны, как отдельные коммерческие решения (класс Custom). Для России это означает:

- разработка российских суперкомпьютеров высшей производительности, скорее всего, не сможет обойтись без интерконнекта Custom. А так как его невозможно купить, то, с неизбежностью, его придется разработать самим (задача выполняемая и уже не раз была успешно реализована);
- младшие и средние российские суперкомпьютеры могут использовать для интерконнекта технологии Infiniband и Ethernet.

Из разработок интерконнекта в России можно отметить:

- ИПС имени А.К. Айламазяна РАН – российские интерконнекты SKIF 3D-torus и «Паутина»;
- НИЦЭВТ – Ангара;
- ВНИИЭФ – СМПО;
- НИИ «Квант» и ИПМ им. М.В. Келдыша РАН – МВС-Экспресс.

Отметим, что единственный отечественный интерконнект, который был не только разработан, но и был использован в системе, вошедшей в рейтинг Top500 – это SKIF 3D-torus (ИПС имени А.К. Айламазяна РАН).

Для построения систем высшей производительности перспективными центральными процессорами являются либо топовые (значит, ограниченно доступные) модели самых успешных коммерческих разработок – сегодня это процессоры AMD – либо собственные процессоры, на базе одной из доступных архитектур, таких как ARM (доказал свою перспективность для такого использования в суперкомпьютерах) и, возможно, RISC-V.

Среди компаний-производителей компьютеров произошли существенные изменения в последние годы. Китайские производители осуществили серьезную интервенцию в отрасли: на их долю в ноябре 2019 года приходилось 63.0% количества сделок и 43.4% объемов (доли $\Sigma R_{\max}$).

Больше всего доля Китая среди систем начального уровня. Для построения систем высшей производительности важен следующий факт – такие системы:

- всегда используются для исследований и разработок; для генерации новых знаний, технологий, материалов; для научных расчётов;
- никогда не используются напрямую для задач реальной экономики: инженерных расчетов, финансовых приложений, услуг связи, интернет-сервисов и т. п.

Это накладывает свой отпечаток на класс решаемых задач: на самом переднем крае научного поиска за время жизни рекордной установки (около 5–7 лет) могут возникнуть особые задачи, требующие для решения особые математические и алгоритмические инструменты. Это следует учитывать при разработке архитектурных решений в части и аппаратных, и программных средств таких систем.

На сегодня США, Китай, Япония и Евросоюз выстроили добротную СК-инфраструктуру, как базис перехода к цифровой экономике, имеют весомые доли своих стран в совокупной мировой производительности суперкомпьютеров. И доли эти больше, чем доли этих стран в мировом ВВП (исключение – Китай, который скрывает свои успехи в СКТ). Такое соотношение хорошо иллюстрирует реальное продвижение стран на пути к цифровой экономике. Россия на сегодня не имеет никакой осмысленной СК-инфраструктуры, а ее доля в совокупной мировой производительности суперкомпьютеров меньше ее доли в мировом ВВП. В последние годы в области СКТ неуклонно растет отставание России от других ведущих стран и отставание России от мирового уровня развития СКТ.

Все эти обстоятельства следует самым серьезным образом учитывать при формировании государственной политики в области СКТ и цифровых технологий.

Список литературы

- [1] Ezell S. J., Atkinson R. D. *The vital importance of high-performance computing to U.S. competitiveness.* – Washington, DC, USA: ITIF. – 28.04.2016. – 58 pp.  [↑50](#)
- [2] Lucas R., Ang J., Bergman K., Borkar Sh., Carlson W., Carrington L., Chiu G., Colwell R., Dally W., Dongarra J., et al. *Top ten exascale research challenges*, DOE advanced scientific computing advisory subcommittee (ASCAC) report. – United States: USDOE Office of Science (SC). – 2014. – 86 pp.  [↑50](#)
- [3] Абрамов С. М. *Правда, искажающая истину. Как следует анализировать Top500?* Вестник ЮУрГУ. Сер. Выч. матем. информ. – 2013. – Т. 2. – № 3. – С. 5–31.   [↑50](#)

- [4] Абрамов С. М. *Июнь 2019: анализ развития суперкомпьютерной отрасли в России и в мире* // Программные системы: теория и приложения.– 2019.– Т. **10**.– № 3.– С. 3–40.   ↑₅₀
- [5] Kogge P. M., Dally W. J. *Frontier vs the Exascale Report: Why so long? and Are We Really There Yet? 2022 IEEE/ACM International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)* (13–18 November 2022, Dallas, TX, USA).– IEEE.– 2023.– Pp. 26–35.  ↑₅₁
- [6] Zheng W. *Research trend of large-scale supercomputers and applications from the TOP500 and Gordon Bell Prize* // Science China Information Sciences.– 2020.– Vol. **63**.– id. 171001.  ↑₅₁
- [7] Khan A., Sim H., Vazhkuudai S. S., Butt A. R., Kim Y. *An analysis of system balance and architectural trends based on Top500 supercomputers* // *The International Conference on High Performance Computing in Asia-Pacific Region, HPC Asia 2021*, New York, NY, USA: Association for Computing Machinery.– 2021.– ISBN 978-1-4503-8842-9.– Pp. 11–22.  ↑₅₁
- [8] *Supercomputer “K computer” takes first place in World, Achieves world’s best performance of 8.162 petaflops to lead TOP500 list.*– RIKEN, Fujitsu Ltd.– 2011 (revised as of December 17, 2018).  ↑₅₇
- [9] Brodtkin J. *With 16 petaflops and 1.6M cores, DOE supercomputer is worlds fastest. Giant cluster built to test nuclear weapons stockpile breaks HPC speed record.*– Ars Technica.– 20.06.2011.  ↑₅₇
- [10] Oak Ridge Leadership Computing Facility *ORNL debuts titan supercomputer.*– Oak Ridge National Laboratory.– 26.10.2012.  ↑₅₇
- [11] Alba D. *Chinas Tianhe-2 Caps Top 10 Supercomputers*, China’s Tianhe-2 outflanks U.S. systems to take No. 1 spot in Top500 list // IEEE Spectrum.– 17.06.2013.  ↑₅₇
- [12] Dongarra J. *Report on the Sunway TaihuLight*, Tech Report UT-EECS-16-742.– Oak Ridge National Laboratory.– 14.06.2016.– 24 pp.  ↑₅₇
- [13] Oak Ridge National Laboratory *Introducing the Summit Supercomputer.*– ORNL, IBM, NVIDIA.– 2020.– 24 pp. ↑₅₇
- [14] Vergara L. V. G., Joubert W., Brim M. J., Budiardja R. D., Maxwell D., Ezell M., Zimmer C., Boehm S., Elwasif W., Oral S., Fuson C., Pelfrey D., Hernandez O., Leverman D., Hanley J., Berrill M., Tharrington A. *Scaling the Summit: deploying the World’s fastest supercomputer* // *High Performance Computing* (Frankfurt, Germany, June 16–20, 2019), Lecture Notes in Computer Science.– vol. **11887**, eds. Weiland M., Juckeland G., Alam S., Jagode H., Cham: Springer.– 2019.– ISBN 978-3-030-34355-2.– Pp. 330–351.  ↑₅₇
- [15] Dongarra J. *Report on the Fujitsu Fugaku system*, Tech Report ICLUT-20-06.– University of Tennessee-Knoxville Innovative Computing Laboratory.– 2020.– 18 pp. ↑₅₇

- [16] Coghlan S. *Argonnes Aurora Exascale Computer*, Smoky Mountain Computational Sciences And Engineering Conference SMC 2019 (August 29, 2019).– 17 pp.  [↑62](#)
- [17] Grider G. *Update on ASC Crossroads System. History and Plans*, From LA-UR-19-24930.– Los alamos national laboratory.– 2019.– 11 pp.  [↑73](#)
- [18] Prickett N. H. *China has already reached exascale — on two separate systems.*– The Next Platform.– October 26, 2021.  [↑80](#)
- [19] *The global economy by GDP.*– BusinessLIVE.– March 13, 2021.  [↑85](#)

Поступила в редакцию 18.04.2023;
 одобрена после рецензирования 04.05.2023;
 принята к публикации 20.05.2023;
 опубликована онлайн 07.07.2023.

Рекомендовал к публикации

Анд. В. Климов

Информация об авторах:

Николай Сергеевич Абрамов

к.т.н., вед. научн. сотр. Института программных систем имени А. К. Айламазяна РАН. Научные интересы: компьютерная графика, распознавание образов, обработка изображений, нейронные сети, искусственный интеллект, принятие решений, управление видеокамерой, геометрия.



0000-0002-1612-3879

e-mail: n-say@nsa.pereslavl.ru

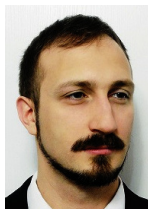


Foto by N. N. Abramova, CC-BY-SA

Сергей Михайлович Абрамов

д.ф.-м.н., чл.-корр. РАН, гл. научн. сотр. Института программных систем имени А. К. Айламазяна РАН. Научные интересы: супер-компьютерные технологии, метавычисления, сетевые технологии, интернет вещей.



0000-0001-6603-7971

e-mail: abram@botik.ru



Foto by A. Yu. Fomenko, CC-BY-SA

Вклад авторов: Н. С. Абрамов – 50% (идея, методология, сбор материала, доработка и редактирование); С. М. Абрамов – 50% (идея, методология, анализ списка Top500, иллюстрации и таблицы, визуализация).

Авторы заявляют об отсутствии конфликта интересов.



June 2022 Top500 List Overview

Nikolai Sergeevich **Abramov**¹, Sergei Mikhailovich **Abramov**²

Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

¹n-say@nsa.pereslavl.ru

Abstract. The article is devoted to the analysis of the current state of the supercomputer industry in the world and in Russia and the prospects for its development. In terms of methodological approach and tools, this article is a continuation of a series of similar analytical reviews by the authors. The novelty of this work is not only in updating the information, taking into account the latest editions of the Top500 list, but also in focusing on the following: the global supercomputer industry is undergoing a radical restructuring—the transition from the «petascale era» to the «exascale era». Russia will have to make the same transition with some delay. Therefore, it is important to understand technological trends and features of solutions for the most powerful systems in the world in recent years. This circumstance is given special attention in the work. And such an emphasis on trend analysis is one of the motivational goals of this work. (*In Russian*).

Key words and phrases: Top500, supercomputers, hybrid architectures, interconnect

2020 *Mathematics Subject Classification:* 65Y05; 68M20, 62Q05

For citation: Nikolai S. Abramov, Sergei M. Abramov. *June 2022 Top500 List Overview*. Program Systems: Theory and Applications, 2023, **14**:2(57), pp. 49–93. (*In Russ.*). https://psta.pstiras.ru/read/psta2023_2_49-93.pdf

References

- [1] S. J. Ezell, R. D. Atkinson. *The vital importance of high-performance computing to U.S. competitiveness*, ITIF, Washington, DC, USA, 28.04.2016, 58 pp. 
- [2] R. Lucas, J. Ang, K. Bergman, Sh. Borkar, W. Carlson, L. Carrington, G. Chiu, R. Colwell, W. Dally, J. Dongarra, al. et. *Top ten exascale research challenges*, DOE advanced scientific computing advisory subcommittee (ASCAC) report, USDOE Office of Science (SC), United States, 2014, 86 pp. 
- [3] S. M. Abramov. True judgments that distort the real truth. How to analyze the Top500? *Vestnik YuUrGU. Ser. Vych. matem. inform.*, **2**:3 (2013), pp. 5–31 (in Russian).  
- [4] S. M. Abramov. “June 2019: Analysis of the Development of the Supercomputer Industry in Russia and Worldwide”, *Program Systems: Theory and Applications*, **10**:3 (2019), pp. 3–40 (in Russian). 
- [5] P. M. Kogge, W. J. Dally. Frontier vs the Exascale Report: Why so long? and Are We Really There Yet? *2022 IEEE/ACM International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)* (13–18 November 2022, Dallas, TX, USA), IEEE, 2023, pp. 26–35. 
- [6] W. Zheng. “Research trend of large-scale supercomputers and applications from the TOP500 and Gordon Bell Prize”, *Science China Information Sciences*, **63** (2020), id. 171001. 
- [7] A. Khan, H. Sim, S. S. Vazhkudai, A. R. Butt, Y. Kim. “An analysis of system balance and architectural trends based on Top500 supercomputers”, *The International Conference on High Performance Computing in Asia-Pacific Region, HPC Asia 2021*, Association for Computing Machinery, New York, NY, USA, 2021, ISBN 978-1-4503-8842-9, pp. 11–22. 
- [8] *Supercomputer “K computer” takes first place in World*, Achieves world’s best performance of 8.162 petaflops to lead TOP500 list, RIKEN, Fujitsu Ltd., 2011 (revised as of December 17, 2018). 
- [9] J. Brodtkin. *With 16 petaflops and 1.6M cores, DOE supercomputer is worlds fastest. Giant cluster built to test nuclear weapons stockpile breaks HPC speed record*, *Ars Technica*, 20.06.2011. 
- [10] Oak Ridge Leadership Computing Facility. *ORNL debuts titan supercomputer*, Oak Ridge National Laboratory, 26.10.2012. 
- [11] D. Alba. “Chinas Tianhe-2 Caps Top 10 Supercomputers”, China’s Tianhe-2 outflanks U.S. systems to take No. 1 spot in Top500 list, *IEEE Spectrum*, 17.06.2013. 
- [12] J. Dongarra. *Report on the Sunway TaihuLight*, Tech Report UT-EECS-16-742, Oak Ridge National Laboratory, 14.06.2016, 24 pp. 

- [13] Ridge National Laboratory Oak. *Introducing the Summit Supercomputer*, ORNL, IBM, NVIDIA, 2020, 24 pp.
- [14] L. V. G. Vergara, W. Joubert, M. J. Brim, R. D. Budiardja, D. Maxwell, M. Ezell, C. Zimmer, S. Boehm, W. Elwasif, S. Oral, C. Fuson, D. Pelfrey, O. Hernandez, D. Leverman, J. Hanley, M. Berrill, A. Tharrington. “Scaling the Summit: deploying the World’s fastest supercomputer”, *High Performance Computing* (Frankfurt, Germany, June 16–20, 2019), Lecture Notes in Computer Science, vol. **11887**, eds. Weiland M., Juckeland G., Alam S., Jagode H., Springer, Cham, 2019, ISBN 978-3-030-34355-2, pp. 330–351. 
- [15] J. Dongarra. *Report on the Fujitsu Fugaku system*, Tech Report ICLUT-20-06, University of Tennessee-Knoxville Innovative Computing Laboratory, 2020, 18 pp.
- [16] S. Coghlan. *Argonnes Aurora Exascale Computer*, Smoky Mountain Computational Sciences And Engineering Conference SMC 2019 (August 29, 2019), 17 pp. 
- [17] G. Grider. *Update on ASC Crossroads System. History and Plans*, From LA-UR-19-24930, Los alamos national laboratory, 2019, 11 pp. 
- [18] N. H. Prickett. *China has already reached exascale — on two separate systems*, The Next Platform, October 26, 2021. 
- [19] *The global economy by GDP*, BusinessLIVE, March 13, 2021. 