Производительность параллельного вывода данных в NetCDF в модели Земной системы ИВМ РАН

М. А. Тарасевич[1,2,3], И. В. Цыбулин[4], В В. Брагина[1,2], Е. М. Володин[1,2]

[1]Институт вычислительной математики им. Г. И. Марчука РАН

[2]Гидрометеорологический научно-исследовательский центр РФ

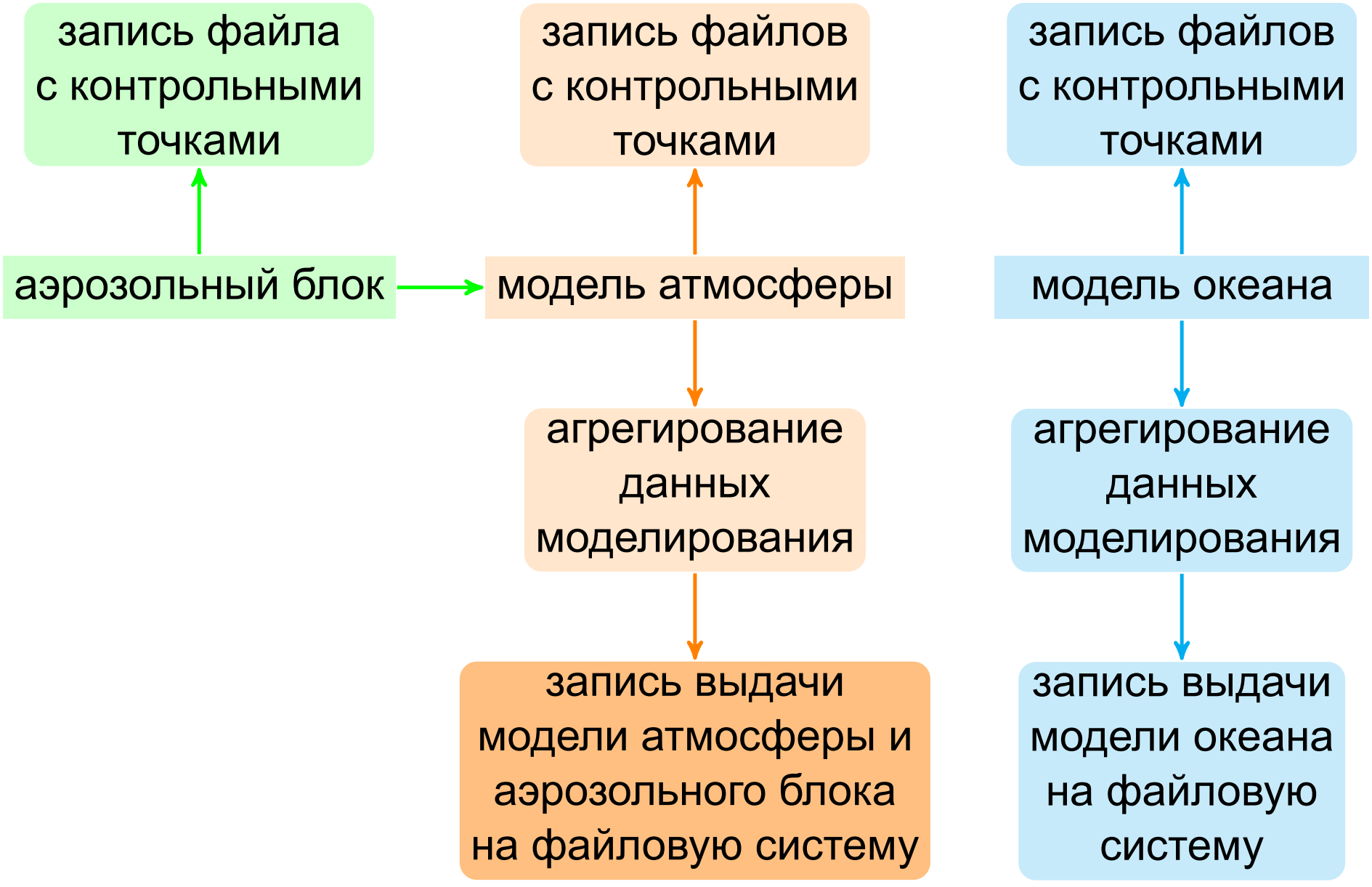
[3]Московский физико-технический институт

[4]Яндекс.Технологии

mashatarasevich@gmail.com

23 сентября 2024 г.

Вывод модели Земной системы ИВМ РАН



Выдача модели атмосферы

- Наиболее влияющая на производительность записи на файловую систему выдача

Тип	Количество полей	Периодичность	Размерность поля
dxy	71	раз в сутки	$nlat \times nlon$
dxyz	7	раз в сутки	$nplevs \times nlat \times nlon$
dxys	10	раз в сутки	$nslevs \times nlat \times nlon$
6h	44	раз в 6 часов	$nlat \times nlon$
3h	37	раз в 3 часа	$nlat \times nlon$
1h	2	раз в час	$nlat \times nlon$

- Для INMCM6P $nlat = 120$, $nlon = 180$, $nplevs = 17$, $nslevs = 21$, $dxy + dxyz + dxys + 6h + 3h + 1h = 79.5$ МБ в сутки
- Для INMCM6M $nlat = 180$, $nlon = 288$, $nplevs = 26$, $nslevs = 73$, $dxy + dxyz + dxys + 6h + 3h + 1h = 312$ МБ в сутки

Запись выдачи модели атмосферы

- Все поля агрегируются в распределённых массивах
- Для записи в файл выполняется сборка распределённых массивов на мастер-процессе модели атмосферы
- Мастер-процесс записывает собранный массив на файловую систему
- Каждое поле dxy, dxyz, dxys, 6h, 3h, 1h выдачи пишется в отдельный неформатный файл прямого доступа
- Остальная выдача (среднесуточная осредненная по долготе и вся среднемесячная) пишется в неформатные файлы прямого доступа с группировкой по типу выдачи

Проблемы вывода модели атмосферы

- При записи неформатных файлов прямого доступа не предусмотрена запись метаданных
- Для расчёта на 1 месяц с записью dxy, dxyz, dxys, 6h, 3h, 1h выдачи доля времени на вывод приведена в таблице

INMCM6P	Модель атмосферы, MPI	14	32	56	72	90	120	180	240	320
	Доля вывода, Cray X40-LC, %	3	5	9	10	11	12	15	19	16
	Доля вывода, кластер ИВМ РАН, %	7	12	18	21	23	28	33	36	38
INMCM6M	Модель атмосферы, MPI	84	120	144	180	208	240	288	312	360
	Доля вывода, Cray X40-LC, %	2	3	4	4	5	5	7	6	6
	Доля вывода, кластер ИВМ РАН, %	6	8	8	11	11	12	13	15	16

- Для INMCM6P время записи может занимать более 15% времени расчёта, для INMCM6M — 5–15% времени расчёта

Особенности NetCDF

- NetCDF файл содержит метаинформацию и данные
- Метаданные — это размерности полей, их названия и дополнительные атрибуты, хранящиеся в файле
- Редактирование метаданных и запись данных — разные режимы работы с NetCDF файлом
- Библиотека NetCDF позволяет работать с распределенными массивами

Оптимизация работы с файлами NetCDF

- Создание NetCDF файла и редактирование метаданных являются коллективными операциями, требующими синхронизации файловой системы
- По этой причине файл открывается один раз в начале года и закрывается перед началом следующего года
- Запись метаданных для нескольких полей в разные файлы дороже, чем в один файл, так как происходит больше синхронизаций
- Лучше с самого начала записать все метаданные в NetCDF файл и после этого работать только с данными

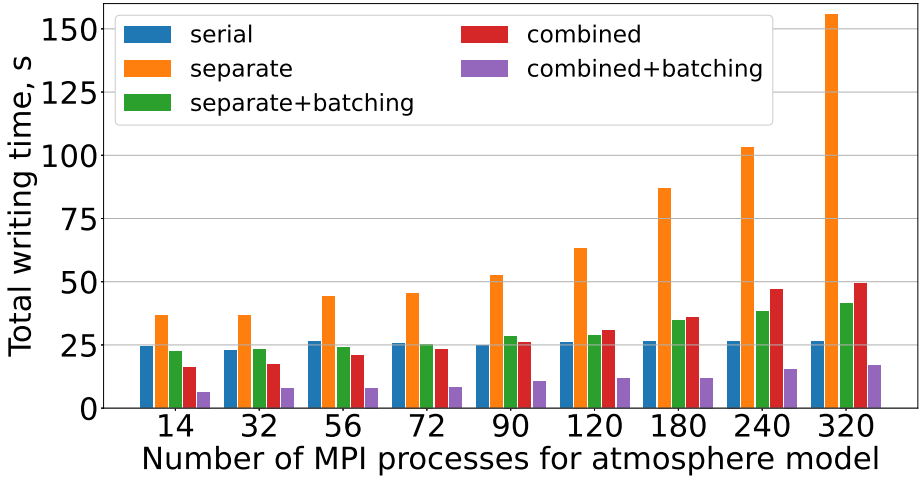
Оптимизация работы с данными в NetCDF

- Необходимо включать коллективный доступ процессов к каждой переменной — без этого каждый процесс пишет на диск неэффективно (мелкая запись, хаотичный доступ)
- Использование неограниченной временной оси приводит к дополнительным затратам на обновление метаданных при записи
- При записи файла, начиная с середины, начало файла по умолчанию заполняется маркером отсутствия данных — на это тратится время и увеличивается потребление места на диске

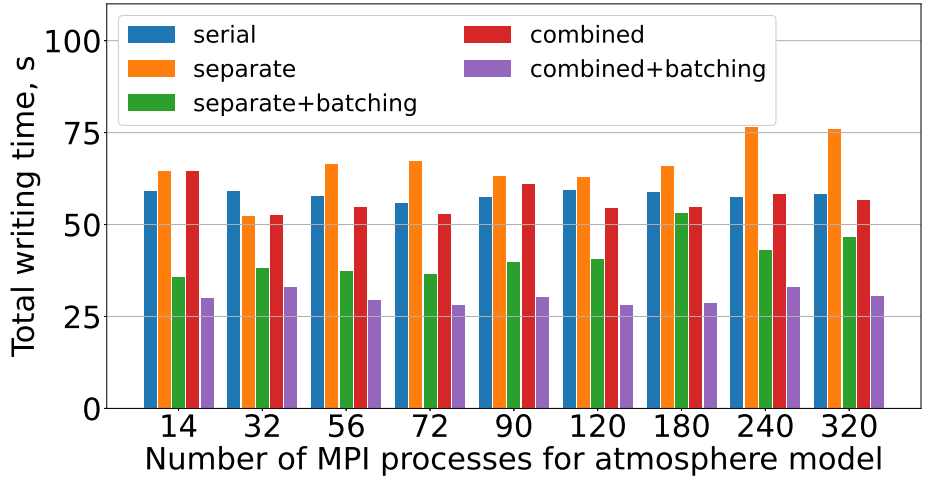
Подходы к записи в NetCDF файлы

- `serial` — исходный способ записи в неформатные файлы прямого доступа
- `separate` — каждое поле помещается в отдельный NetCDF файл, по аналогии с `serial`
- `combined` — запись полей одного типа выдачи в один файл
- `batching` — упаковка («складирование») данных за несколько моментов времени во вспомогательный буфер в памяти перед записью в файл

Производительность записи выдачи

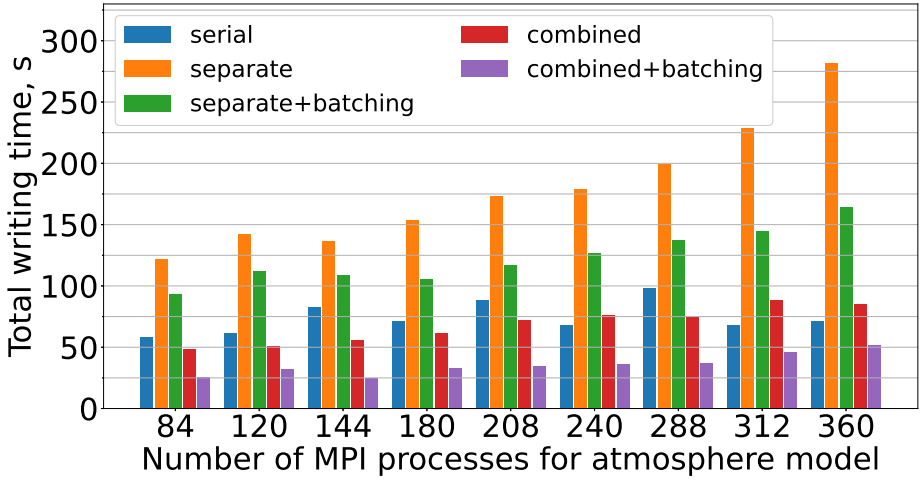


Cray XC40-LC

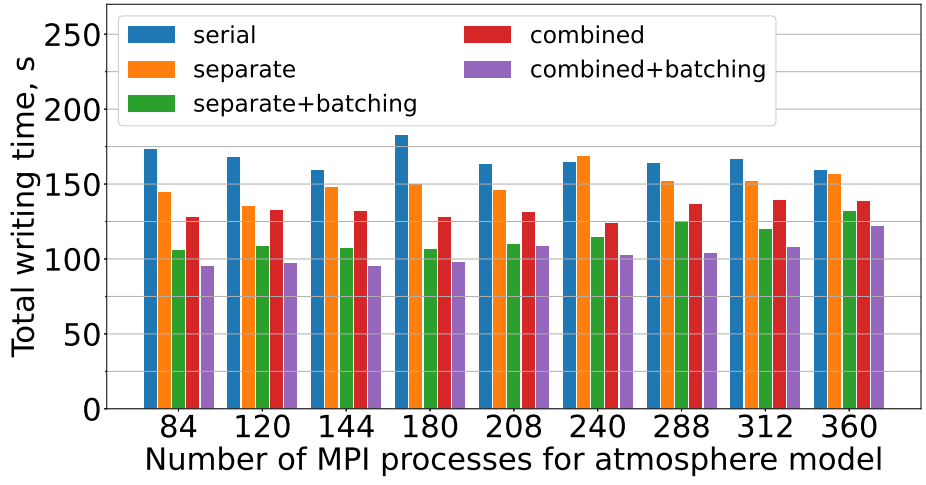


кластер ИВМ РАН

Время вывода для INMCM6P



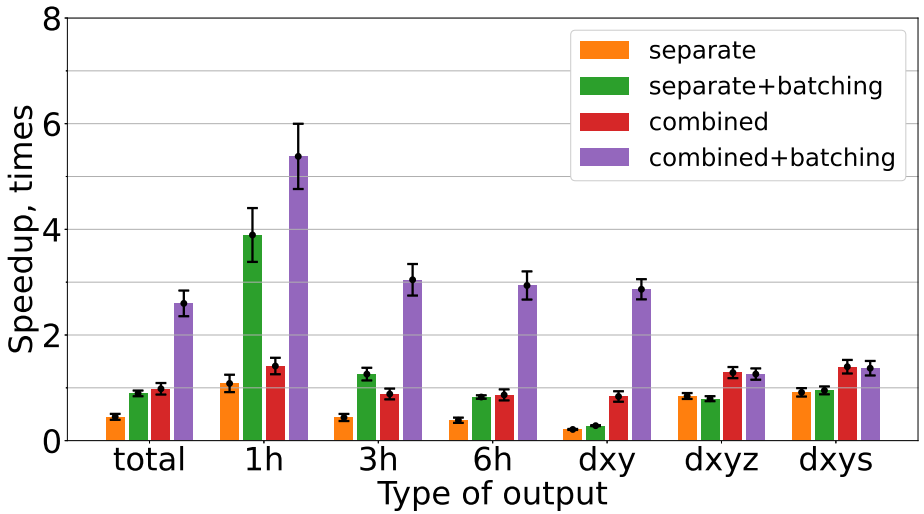
Cray XC40-LC



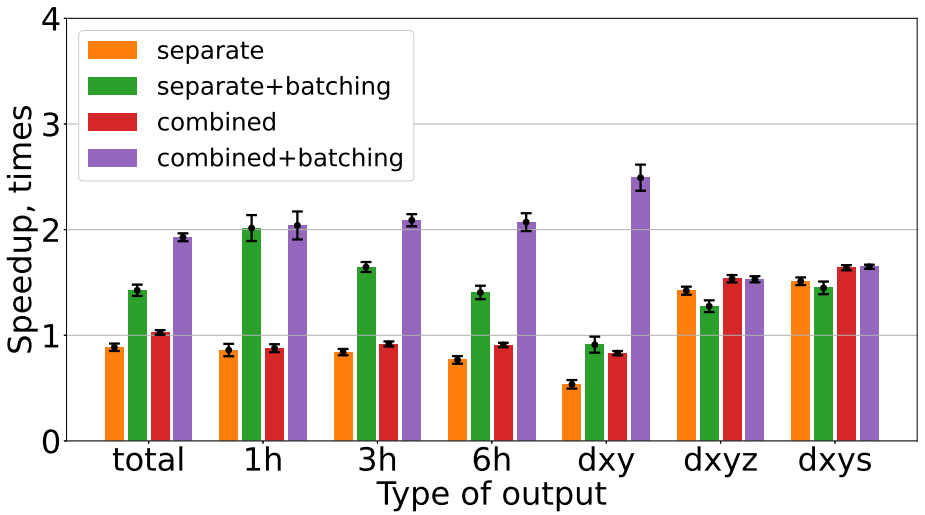
кластер ИВМ РАН

Время вывода для INMCM6M

Ускорение записи различных типов выдачи

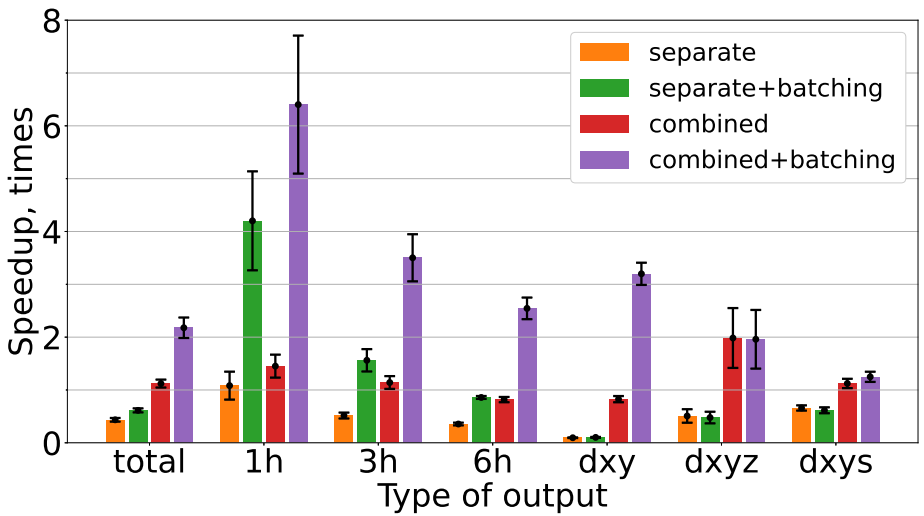


Cray XC40-LC

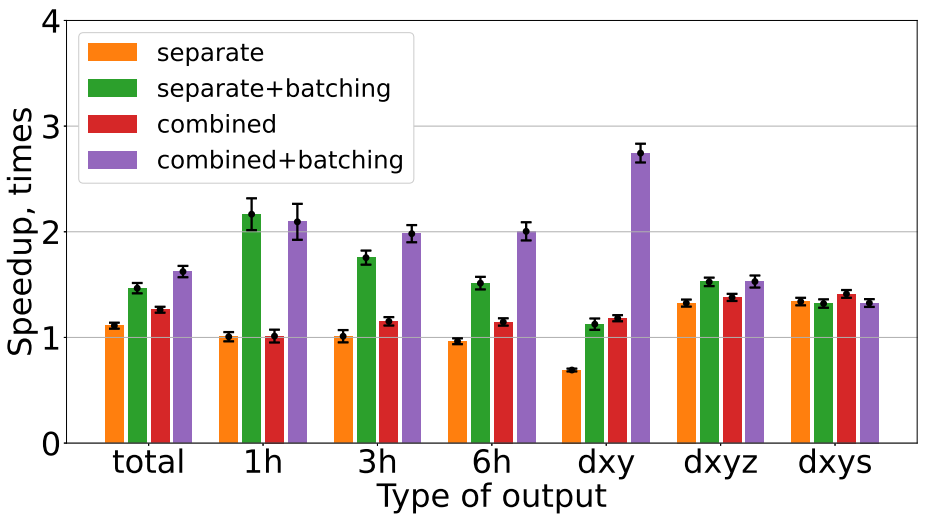


кластер ИВМ РАН

Ускорение записи выдачи для INMCM6P



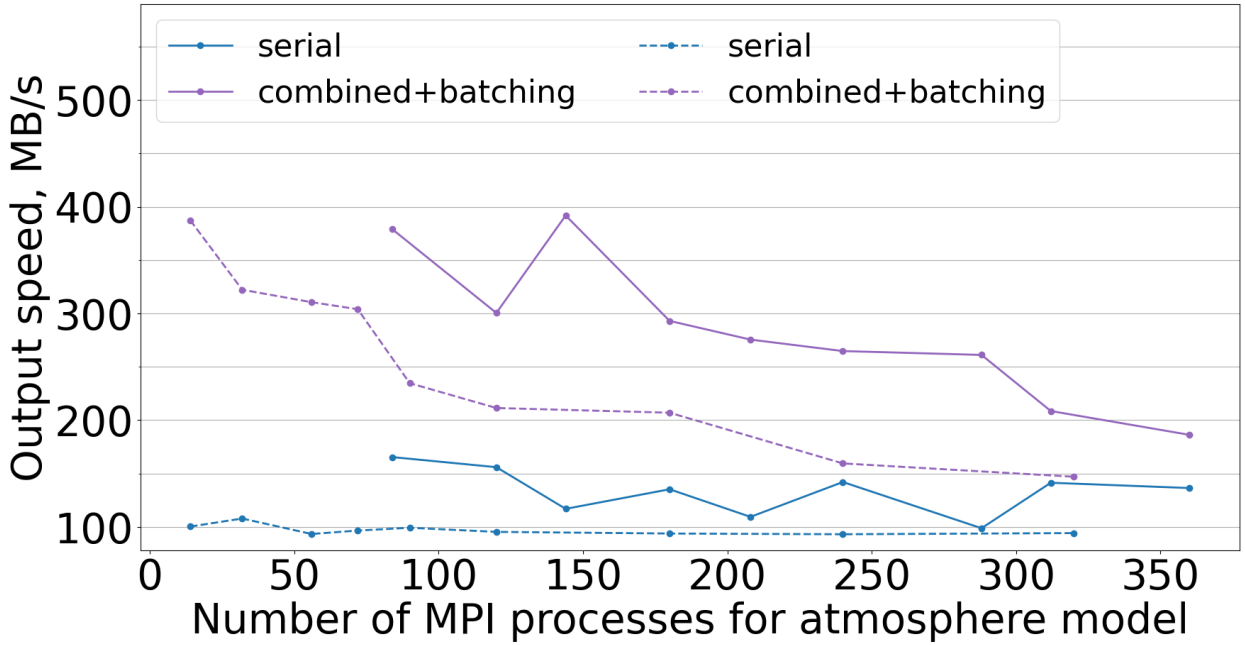
Cray XC40-LC



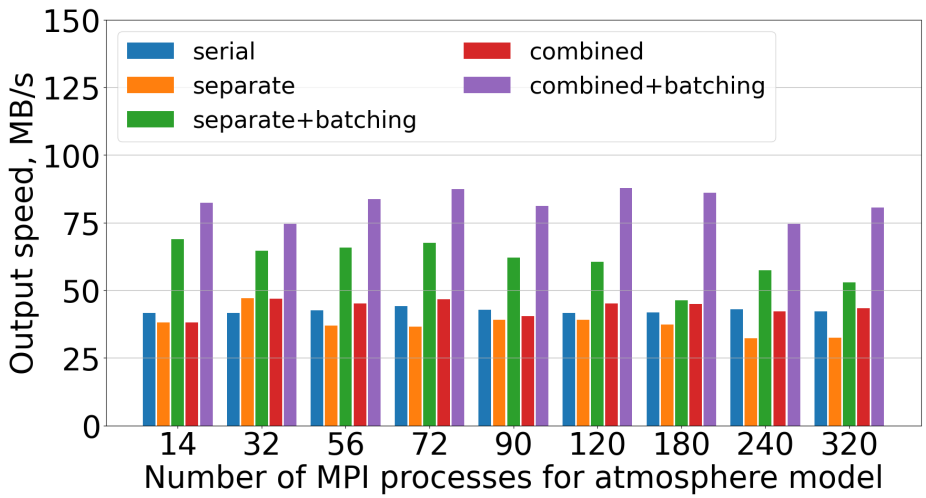
кластер ИВМ РАН

Ускорение записи выдачи для INMCM6M

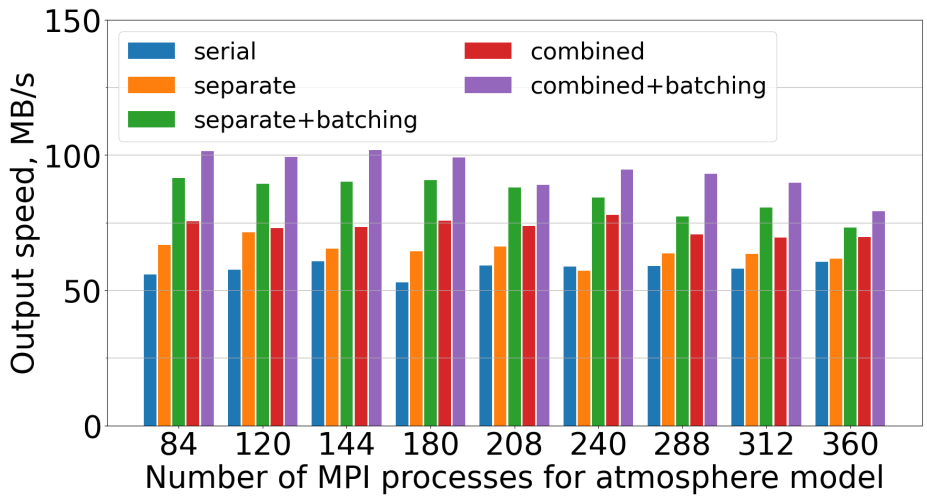
Скорость записи на файловую систему



Cray XC40-LC, Lustre3.2



кластер ИВМ РАН, INMCM6P



кластер ИВМ РАН, INMCM6M

Выводы

- Запись большого количества файлов может значительно снизить быстродействие параллельного вывода в NetCDF
- Для достижения максимальной скорости вывода данные необходимо накапливать перед записью на диск
- На масштабируемость вывода может существенно влиять топология сети
- Хотя вывод данных МЗС и может занимать значимое время работы кода модели, его объемы с учётом большой интенсивности недостаточны, чтобы эффективно использовать распределенную файловую систему

Спасибо за внимание!

Вопросы?

mashatarasevich@gmail.com