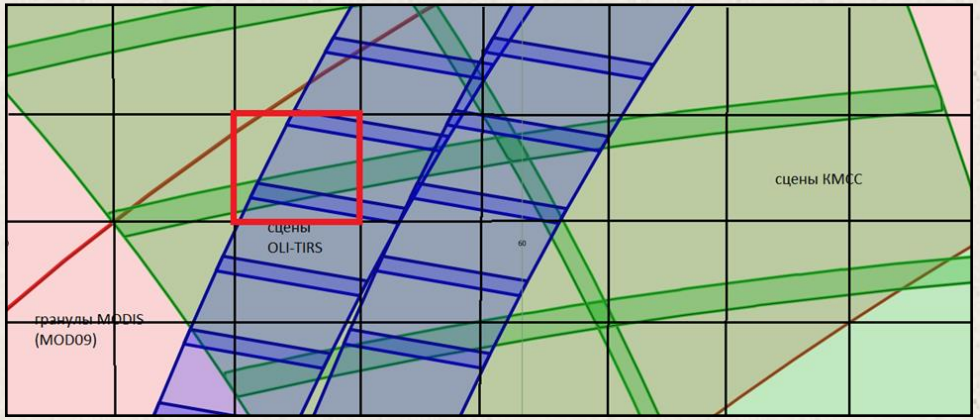
Анализ эффективности системы динамического блочного доступа к данным для предоставления их системам обработки

Прошин А.А., Лупян Е.А.

Институт космических исследований Российской академии наук, Москва

Современные проблемы дистанционного зондирования Земли из космоса, 15-19 ноября 2021г

Многие задачи, связанные с обработкой и визуализацией пространственных данных, могут быть наиболее эффективно решены с использованием блочного подхода, в рамках которого используются данные по тайлам, на которые равномерно разбита вся зона интереса. Это позволяет обеспечить необходимую степень распараллеливания обработки, реализовать гибкий механизм кэширования результатов и снизить нагрузку на централизованные ресурсы. Наиболее распространенным вариантом реализации описываемого подхода является использование заранее подготовленных архивов данных с фиксированным пространственным разбиением. Однако такой вариант реализации оказывается нецелесообразным, когда возникает необходимость использования разных пространственных разбиений для одних и тех же данных, в частности, когда необходимо совместно обрабатывать различные типы спутниковой информации, отличающиеся как по разрешению, так и по организации их хранения в архиве. В таких случаях наиболее эффективным оказывается динамическое формирование блоков данных в таком пространственном разбиении и с такими характеристиками, которые будут оптимальны для решения конкретной задачи по обработке или визуализации данных. С целью реализации такого механизма в ИКИ РАН была разработана технология динамического блочного доступа к архивам спутниковых данных ЦКП «ИКИ-Мониторинг».



Пример покрытия территории контурами фрагментов различных типов спутниковых данных, существенно отличающихся как по размеру, так и по проекции хранения в архивах

Методика оценки временных затрат на подготовку блоков данных

В рамках разработанной технологии динамического блочного доступа к данным подготовка блоков данных производится в параллельном режиме на кластере специализированных серверов, на которых реализован прямой доступ к файлам в архиве. Для того чтобы оценить суммарное время подготовки данных, необходимых для проведения конкретного типа обработки необходимо получить среднее время выполнение запроса на получение одного блока данных.

Время выполнения запроса на подготовку блока данных зависит от множества факторов, таких как производительность систем хранения, скорость сетевых соединений, размер и тип предоставляемых спутниковых данных и др. Установление всех зависимостей без использования упрощенной модели требует чрезмерных времени и ресурсов, что не позволяет оперативно оценить требуемое время для нового типа обработки. Для построения такой модели было проанализировано влияние наиболее значимых из них, и был установлен целый ряд упрощенных зависимостей, которые можно использовать для оценки времени подготовки разных выборок исходных данных. Ниже приводится формула для оценки времени подготовки блока данных по данным одного канала спутникового изображения как сумма зависимостей, которые могут быть установлены экспериментально:

T-channel(n, arch_type, dst_proj, compression) =

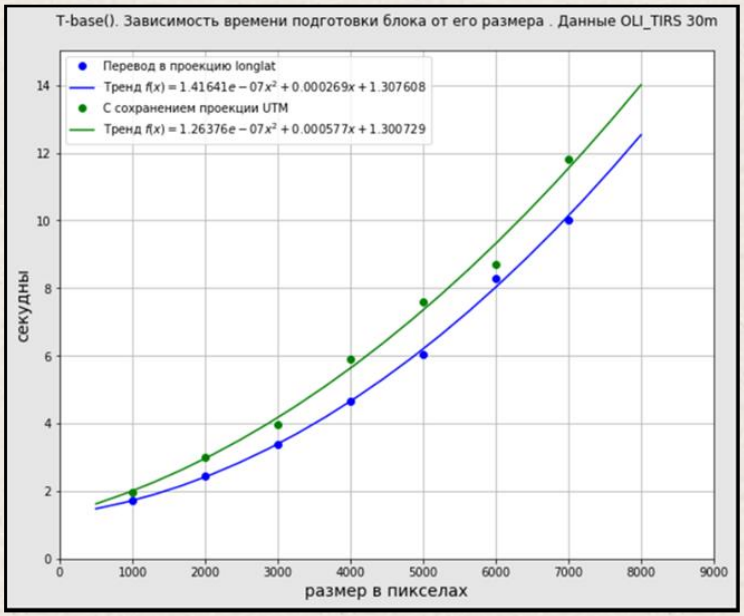
T-base(n, arch_type , dst_proj)	базовое время формирования блока данных
+ T-compression(n, compression)	время сжатия блока данных
+ T-get(size(n, compression))	время передачи блока данных по сети

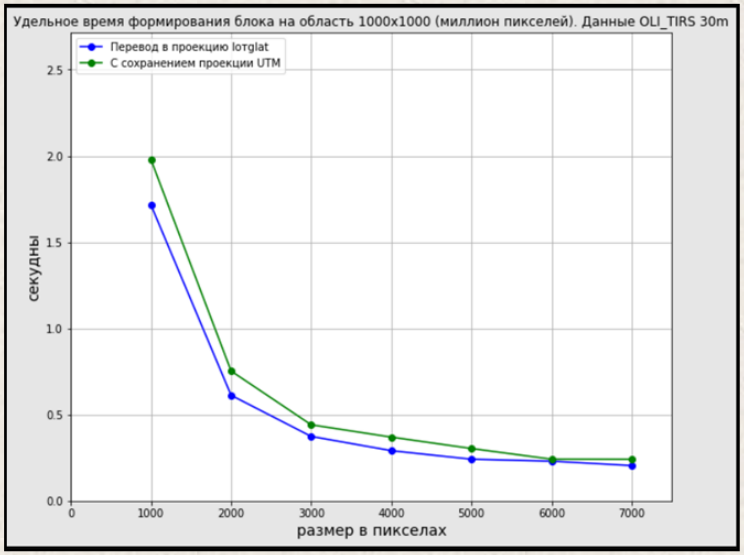
Где **n** – линейный размер блока данных в пикселах,
arch_type – характеристики исходных файлов, включая проекцию и характерный размер фрагментов в архиве
dst_proj - проекция получаемых данных
compression – используемый алгоритм сжатия
size(n, compression) – размер «сжатого» блока данных
Далее кратко приводится методика для получения каждый из вышеперечисленных зависимостей

Получение базовой зависимости T-base

Время формирования блока существенно зависит от производительности и загруженности конкретного узла распределенного архива спутниковых данных. Поэтому для получения базовой зависимости это время усредняется по случайной выборке из полного набора требуемых блоков данных, который необходимо подготовить для проведения конкретной задачи по их обработке. Для того чтобы избежать кэширования результатов, которое может существенно исказить результаты, случайная выборка генерируется для каждого из исследуемых размеров блока в пикселах.

На слайде приводятся примеры базовых зависимостей, полученных при формировании данных по данным прибора OLI_TIRS, установленного на спутниках Landsat, за период с 2013 по 2021 год. Приведены зависимости для случая с сохранением проекции UTM и при переводе в географическую проекцию. На графике видно, что получаемые зависимости хорошо аппроксимируются полиномами второй степени, что, что позволяет на основе небольшого числа замеров для различных размеров блока в пикселах, достаточно точно оценить время подготовки для любого такого размера. Также приведены соответствующие графики для удельного времени подготовки данных. Отметим, что падение эффективности дисковых операций чтения и записи данных при уменьшении объемов данных, вносит существенный вклад практически во все исследуемые зависимости.

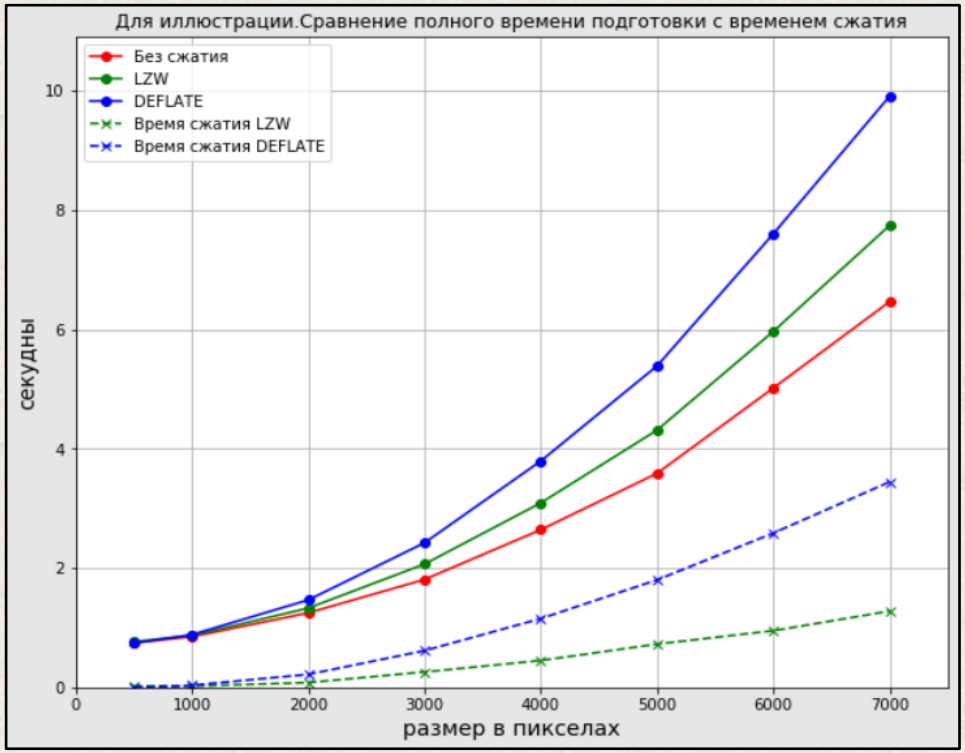




Получение зависимости T-compression

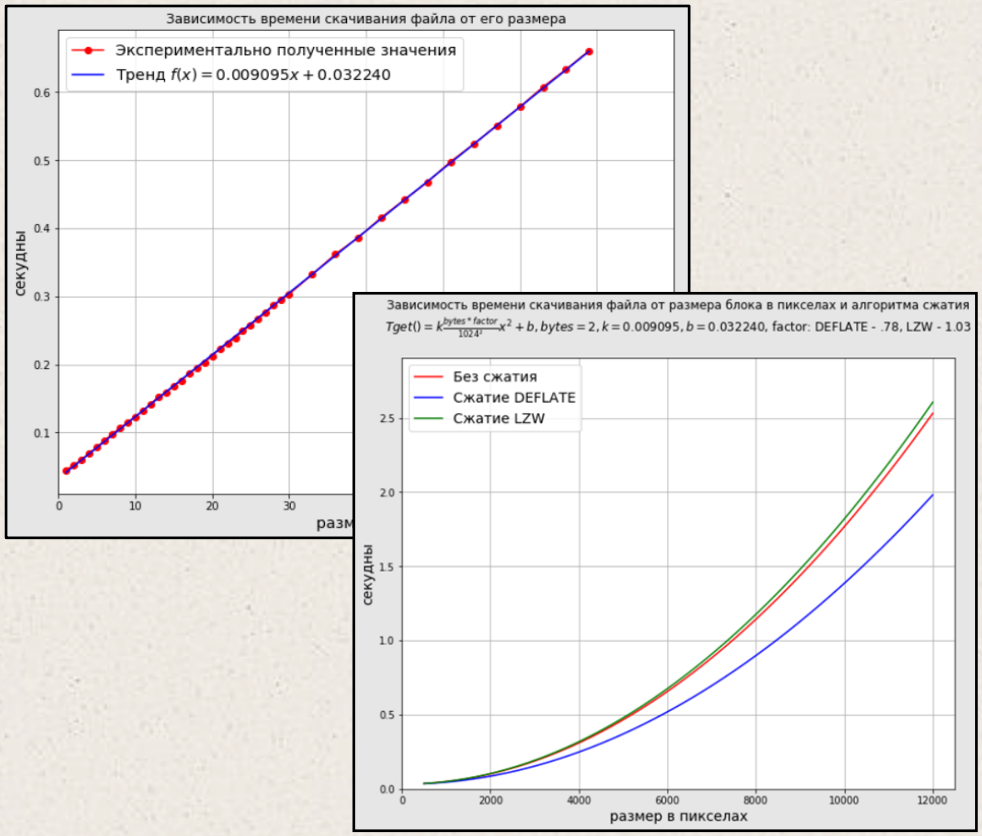
Зависимость времени сжатия от размера блока может быть получается путем вычитания из зависимости времени подготовки блоков с использованием заданного алгоритма сжатия зависимости, полученной для этой процедуры без сжатия. На слайде приведены графики, соответствующие времени подготовки данных без сжатия, с использованием быстрого алгоритма LZW и более эффективного и медленного алгоритма DEFLATE.

Также была оценена степень сжатия данных при использовании каждого из вышеперечисленных алгоритмов. При сжатии полностью заполненных данными блоков процент сжатия для алгоритма DEFLATE стабильно держится в районе 22%, а для алгоритма LZW – размер файлов даже возрастает примерно на 3%. Однако использование алгоритма LZW все-же может быть целесообразным когда значительная часть блоков данных заполнена данными не полностью.



Получение зависимости T-get

Так как время скачивания файлов данных при использовании стандартного протокола HTTP практически не зависит от их содержимого, то сначала можно установить зависимость времени скачивания файлов от их размера в мегабайтах, характерную для имеющейся инфраструктуры. На слайде приведен пример такой зависимости, хорошо аппроксимируемой линейной функцией. На ее основе определяются искомые зависимости времени скачивания блоков данных от их размера в пикселах при использовании различных алгоритмов сжатия, которые аппроксимируются полиномами второй степени с вычисляемыми на основе установленной линейной зависимости коэффициентами.



Заключение

На основании анализа основных факторов, влияющих на эффективность механизма подготовки блоков данных для разных типов исходной информации, была разработана методика, позволяющая оперативно оценивать временные затраты на подготовку заданного набора исходных данных для проведения их обработки. Созданная методика также может быть использована для выбора оптимального размера блока данных и других характеристик предоставляемых данных.

Работа выполнена в рамках темы "Большие данные в космических исследованиях: астрофизика, солнечная система, геосфера" (госрегистрация №0024-2019-0014).

СПАСИБО ЗА ВНИМАНИЕ

ИКИ