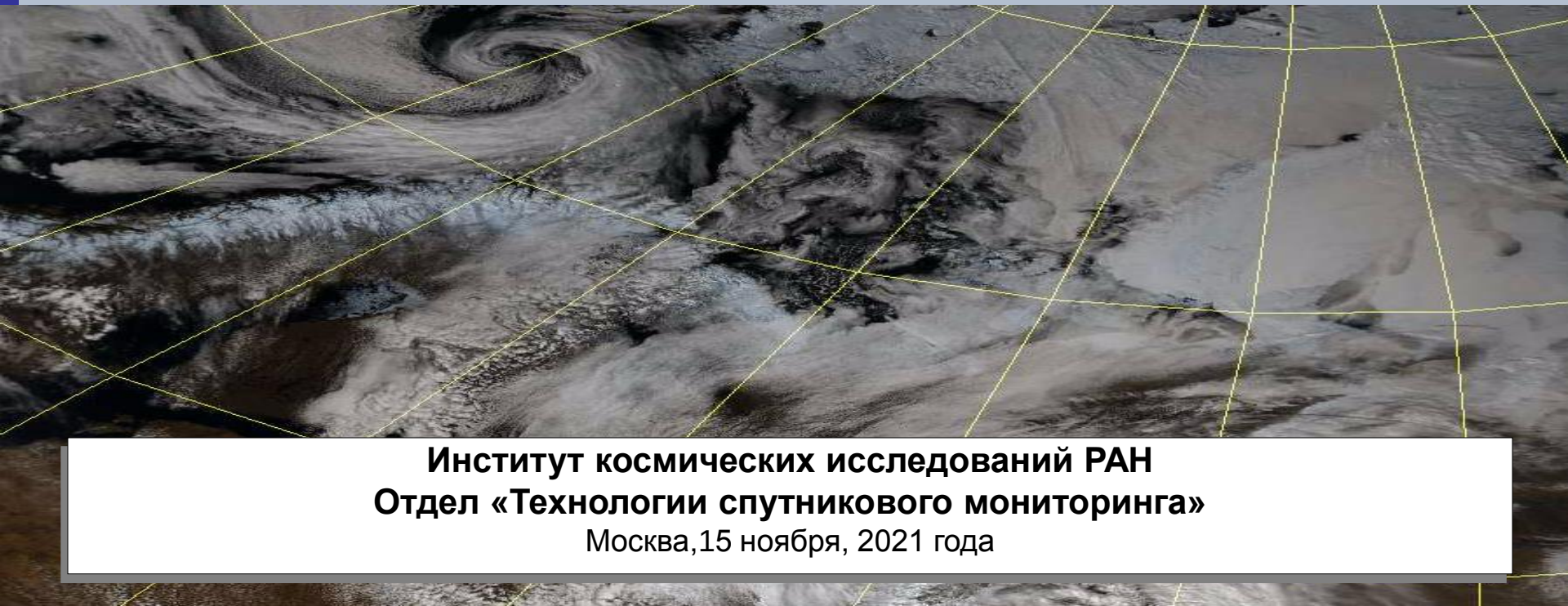
Опыт использования различных подходов к распараллеленной обработке спутниковых данных.

Матвеев А.М., Кобец Д.А., Радченко М.В.



Институт космических исследований РАН
Отдел «Технологии спутникового мониторинга»

Москва,15 ноября, 2021 года

План доклада

1.Общее описание архитектуры системы обработки

2.Контроль загруженности при помощи procs.shtml (простое running/idle)

3.Контроль загруженности обработчиков более детальный (zabbix)

4.BI как инструмент анализа системы обработки

5.Виртуализация, плюсы и минусы. Интегральная скорость выполнения заданий с виртуалками и без.

6.Параллельный запуск нескольких обработок на одной машине

7.Параллельный запуск готовых приложений в рамках 1-й обработки

8. Параллелизм внутри процесса (CreateThread)

9. Планы на будущее

Терминология

Задание – комплект данных под обработку, иногда вместе с софтом обработки (обычно для тяжелых и сверхтяжелых процессов).

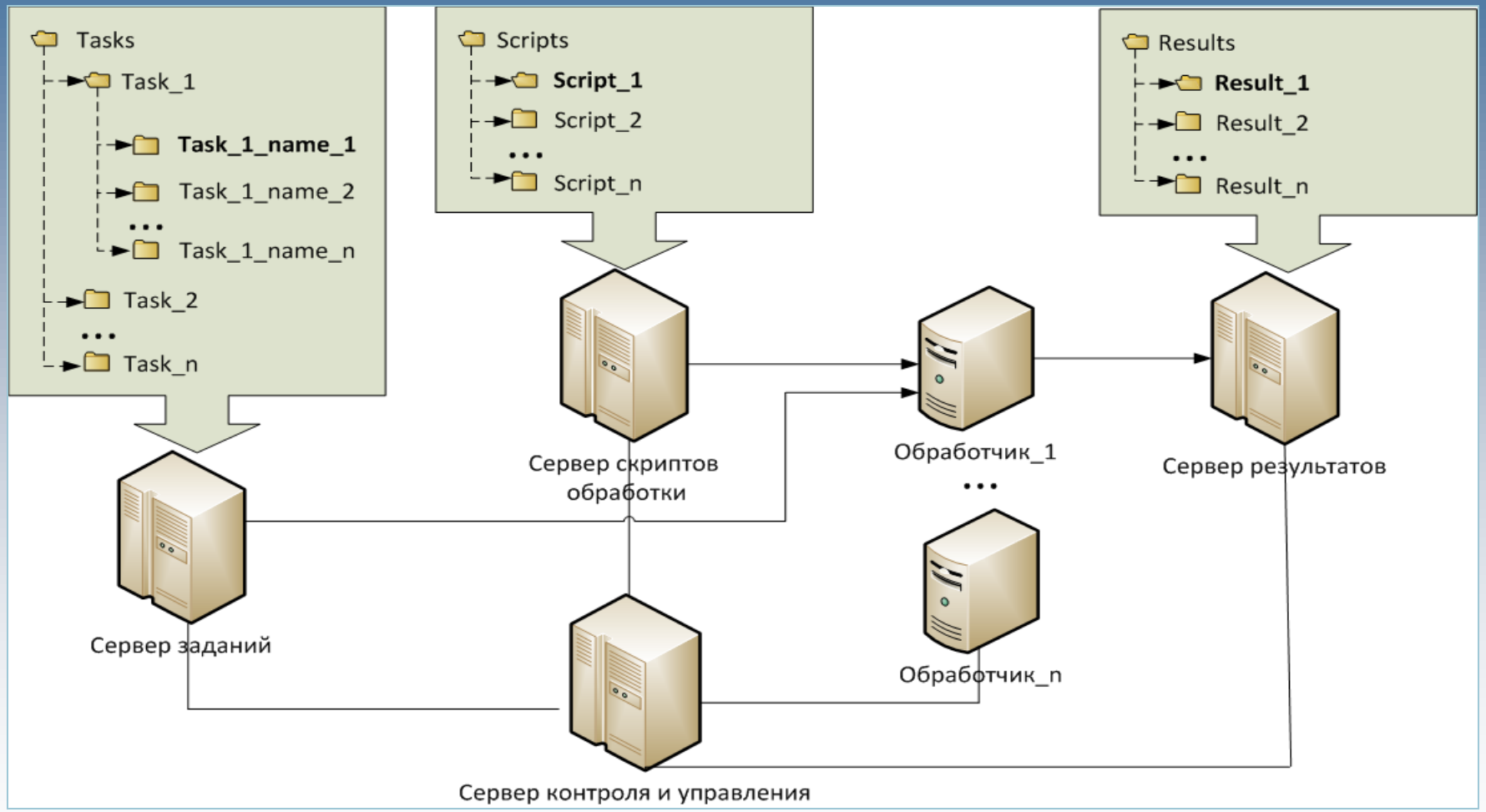
Задача – комплект скриптов и приложений обработки данных.

Обработчик – компьютер для выполнения заданий под управлением windows (есть и юникс-подобные, но в данном докладе не упоминаются).

Обработка – выполнение задания на обработчике

Сервер заданий – UNIX-машина с выделенными директориями, в которых формируются задания, на этом же сервере лежит набор инструкций для обработки задания.

Структура комплекса обработки



Инфраструктура ЦКП Ики-мониторинг

68 обработчиков (из которых 12 виртуальные):
- От 2 (b-виртуалки) до 16 (Blades) ядер
- От 6 (h-виртуалки) до 32 Gb оперативной памяти
- Win7/Win10

~ 200 различных заданий

17 серверов обработки

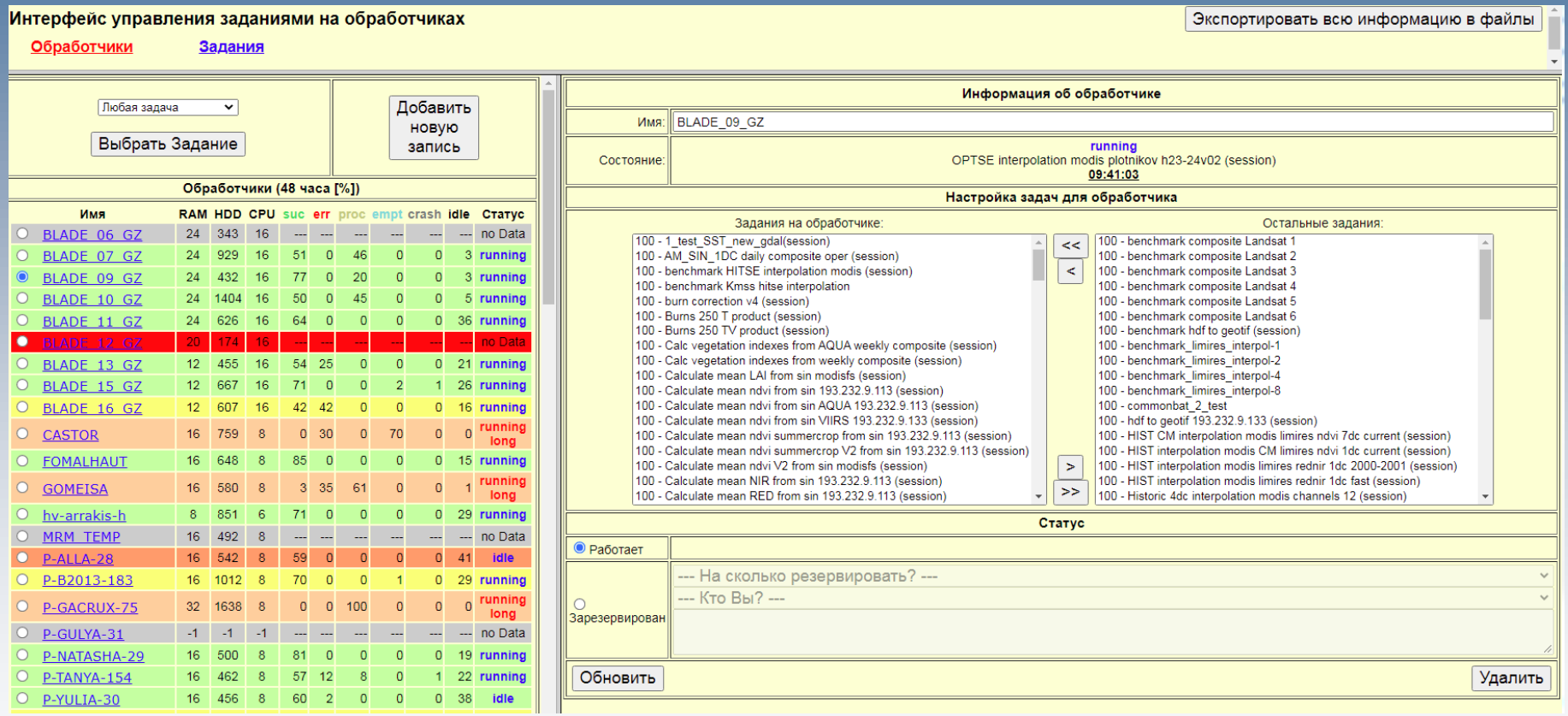
50 серверов хранения – более 5 Петабайт данных

Данные с 37 приборов высокого разрешения

Данные с 25 приборов низкого и среднего разрешения

Софт: ПО Sputnik, Anaconda2-3 (gdal), скрипты – perl/python/bat, web-интерфейс управления настройками обработки, БД обработок - mysql

Интерфейс управления и контроля



Интерфейс управления заданиями на обработчиках

Экспортировать всю информацию в файлы

Обработчики Задания

Любая задача

Выбрать Задание

Добавить новую запись

Обработчики (48 часа [%])

	Имя	RAM	HDD	CPU	suc	err	proc	empt	crash	idle	Статус
○	BLADE_06_GZ	24	343	16	---	---	---	---	---	---	no Data
○	BLADE_07_GZ	24	929	16	51	0	46	0	0	3	running
◉	BLADE_09_GZ	24	432	16	77	0	20	0	0	3	running
○	BLADE_10_GZ	24	1404	16	50	0	45	0	0	5	running
○	BLADE_11_GZ	24	626	16	64	0	0	0	0	36	running
○	BLADE_12_GZ	20	174	16	---	---	---	---	---	---	no Data
○	BLADE_13_GZ	12	455	16	54	25	0	0	0	21	running
○	BLADE_15_GZ	12	667	16	71	0	0	2	1	26	running
○	BLADE_16_GZ	12	607	16	42	42	0	0	0	16	running
○	CASTOR	16	759	8	0	30	0	70	0	0	running long
○	FOMALHAUT	16	648	8	85	0	0	0	0	15	running
○	GOMEISA	16	580	8	3	35	61	0	0	1	running long
○	hv-arrakis-h	8	851	6	71	0	0	0	0	29	running
○	MRM_TEMP	16	492	8	---	---	---	---	---	---	no Data
○	P-ALLA-28	16	542	8	59	0	0	0	0	41	idle
○	P-B2013-183	16	1012	8	70	0	0	1	0	29	running
○	P-GACRUX-75	32	1638	8	0	0	100	0	0	0	running long
○	P-GULYA-31	-1	-1	-1	---	---	---	---	---	---	no Data
○	P-NATASHA-29	16	500	8	81	0	0	0	0	19	running
○	P-TANYA-154	16	462	8	57	12	8	0	1	22	running
○	P-YULIA-30	16	456	8	60	2	0	0	0	38	idle

Информация об обработчике

Имя: BLADE_09_GZ

Состояние: running
OPTSE interpolation modis plotnikov h23-24v02 (session)
09:41:03

Настройка задач для обработчика

Задания на обработчике:

100 - 1_test_SST_new_gdal(session)
100 - AM_SIN_1DC daily composite oper (session)
100 - benchmark HITSE interpolation modis (session)
100 - benchmark Kmss hitse interpolation
100 - burn correction v4 (session)
100 - Burns 250 T product (session)
100 - Burns 250 TV product (session)
100 - Calc vegetation indexes from AQUA weekly composite (session)
100 - Calc vegetation indexes from weekly composite (session)
100 - Calculate mean LAI from sin modisfs (session)
100 - Calculate mean ndvi from sin 193.232.9.113 (session)
100 - Calculate mean ndvi from sin AQUA 193.232.9.113 (session)
100 - Calculate mean ndvi from sin VIIRS 193.232.9.133 (session)
100 - Calculate mean ndvi summercrop from sin 193.232.9.113 (session)
100 - Calculate mean ndvi summercrop V2 from sin 193.232.9.113 (session)
100 - Calculate mean ndvi V2 from sin modisfs (session)
100 - Calculate mean NIR from sin 193.232.9.113 (session)
100 - Calculate mean RED from sin 193.232.9.113 (session)

<< < > >>

Остальные задания:

100 - benchmark composite Landsat 1
100 - benchmark composite Landsat 2
100 - benchmark composite Landsat 3
100 - benchmark composite Landsat 4
100 - benchmark composite Landsat 5
100 - benchmark composite Landsat 6
100 - benchmark hdf to geotif (session)
100 - benchmark_limires_interpol-1
100 - benchmark_limires_interpol-2
100 - benchmark_limires_interpol-4
100 - benchmark_limires_interpol-8
100 - commonbat_2_test
100 - hdf to geotif 193.232.9.133 (session)
100 - HIST CM interpolation modis limires ndvi 7dc current (session)
100 - HIST interpolation modis CM limires ndvi 1dc current (session)
100 - HIST interpolation modis limires rednir 1dc 2000-2001 (session)
100 - HIST interpolation modis limires rednir 1dc fast (session)
100 - Historic 4dc interpolation modis channels 12 (session)

Статус

◉ Работает

○ Зарезервирован

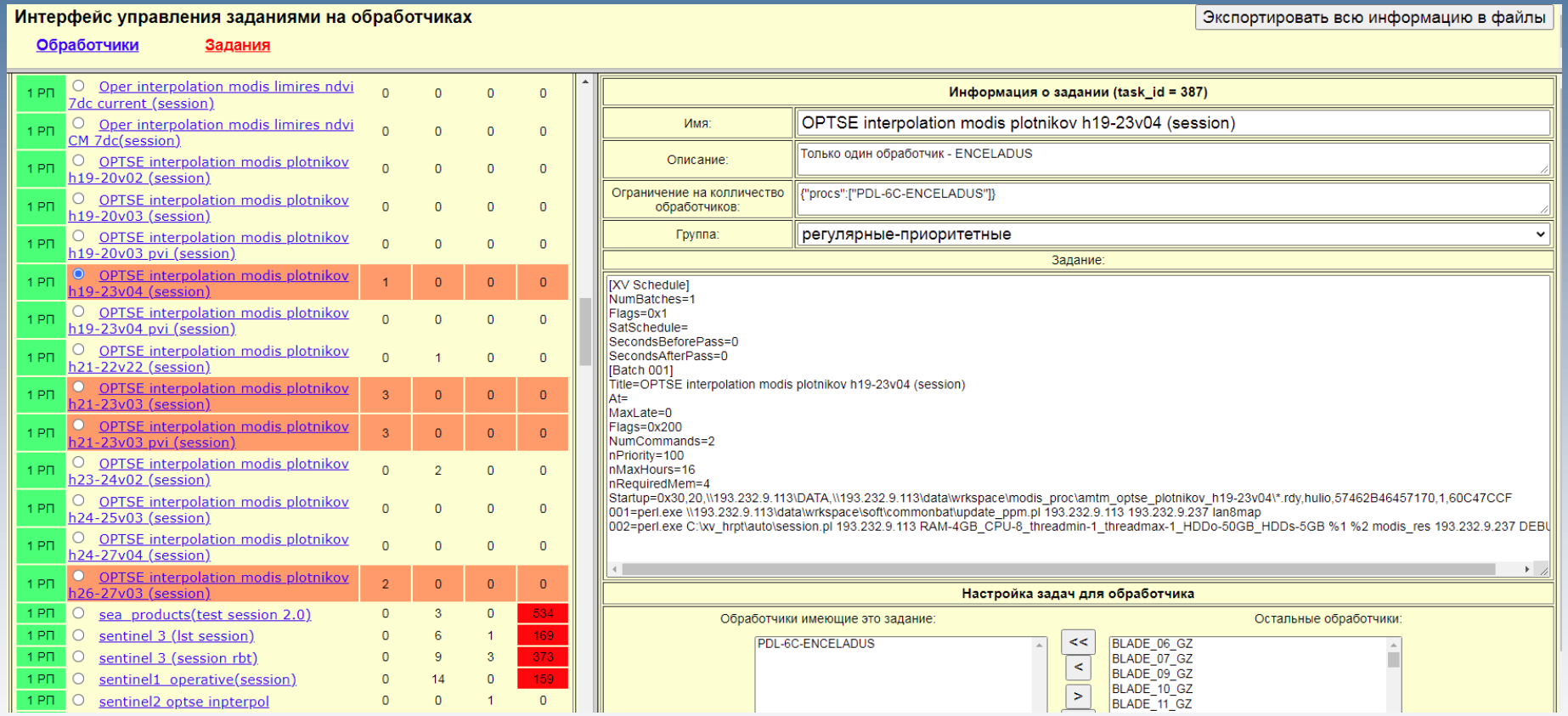
--- На сколько резервировать? ---

--- Кто Вы? ---

Обновить Удалить

ИКИ

Интерфейс управления и контроля



Интерфейс управления заданиями на обработчиках

Экспортировать всю информацию в файлы

Обработчики Задания

1 РП	Oper interpolation modis limires ndvi 7dc current (session)	0	0	0	0
1 РП	Oper interpolation modis limires ndvi CM 7dc(session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h19-20v02 (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h19-20v03 (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h19-20v03 pvi (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h19-23v04 (session)	1	0	0	0
1 РП	OPTSE interpolation modis plotnikov h19-23v04 pvi (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h21-22v22 (session)	0	1	0	0
1 РП	OPTSE interpolation modis plotnikov h21-23v03 (session)	3	0	0	0
1 РП	OPTSE interpolation modis plotnikov h21-23v03 pvi (session)	3	0	0	0
1 РП	OPTSE interpolation modis plotnikov h23-24v02 (session)	0	2	0	0
1 РП	OPTSE interpolation modis plotnikov h24-25v03 (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h24-27v04 (session)	0	0	0	0
1 РП	OPTSE interpolation modis plotnikov h26-27v03 (session)	2	0	0	0
1 РП	sea_products(test session 2.0)	0	3	0	534
1 РП	sentinel 3 (lst session)	0	6	1	169
1 РП	sentinel 3 (session rbt)	0	9	3	373
1 РП	sentinel1_operative(session)	0	14	0	159
1 РП	sentinel2 optse inpterpol	0	0	1	0

Информация о задании (task_id = 387)

Имя:	OPTSE interpolation modis plotnikov h19-23v04 (session)
Описание:	Только один обработчик - ENCELADUS
Ограничение на колличество обработчиков:	{"procs":["PDL-6C-ENCELADUS"]}
Группа:	регулярные-приоритетные

Задание:

```
[XV Schedule]
NumBatches=1
Flags=0x1
SatSchedule=
SecondsBeforePass=0
SecondsAfterPass=0
[Batch 001]
Title=OPTSE interpolation modis plotnikov h19-23v04 (session)
At=
MaxLate=0
Flags=0x200
NumCommands=2
nPriority=100
nMaxHours=16
nRequiredMem=4
Startup=0x30,20,\\193.232.9.113\DATA,\\193.232.9.113\data\wrkspace\modis_proc\amtm_optse_plotnikov_h19-23v04\*.rdy,hulio,57462B46457170,1,60C47CCF
001=perl.exe \\193.232.9.113\data\wrkspace\soft\commonbat\update_ppm.pl 193.232.9.113 193.232.9.237 lan8map
002=perl.exe C:\xv_hrpt\auto\session.pl 193.232.9.113 RAM-4GB_CPU-8_threadmin-1_threadmax-1_HDDo-50GB_HDDs-5GB %1 %2 modis_res 193.232.9.237 DEBU
```

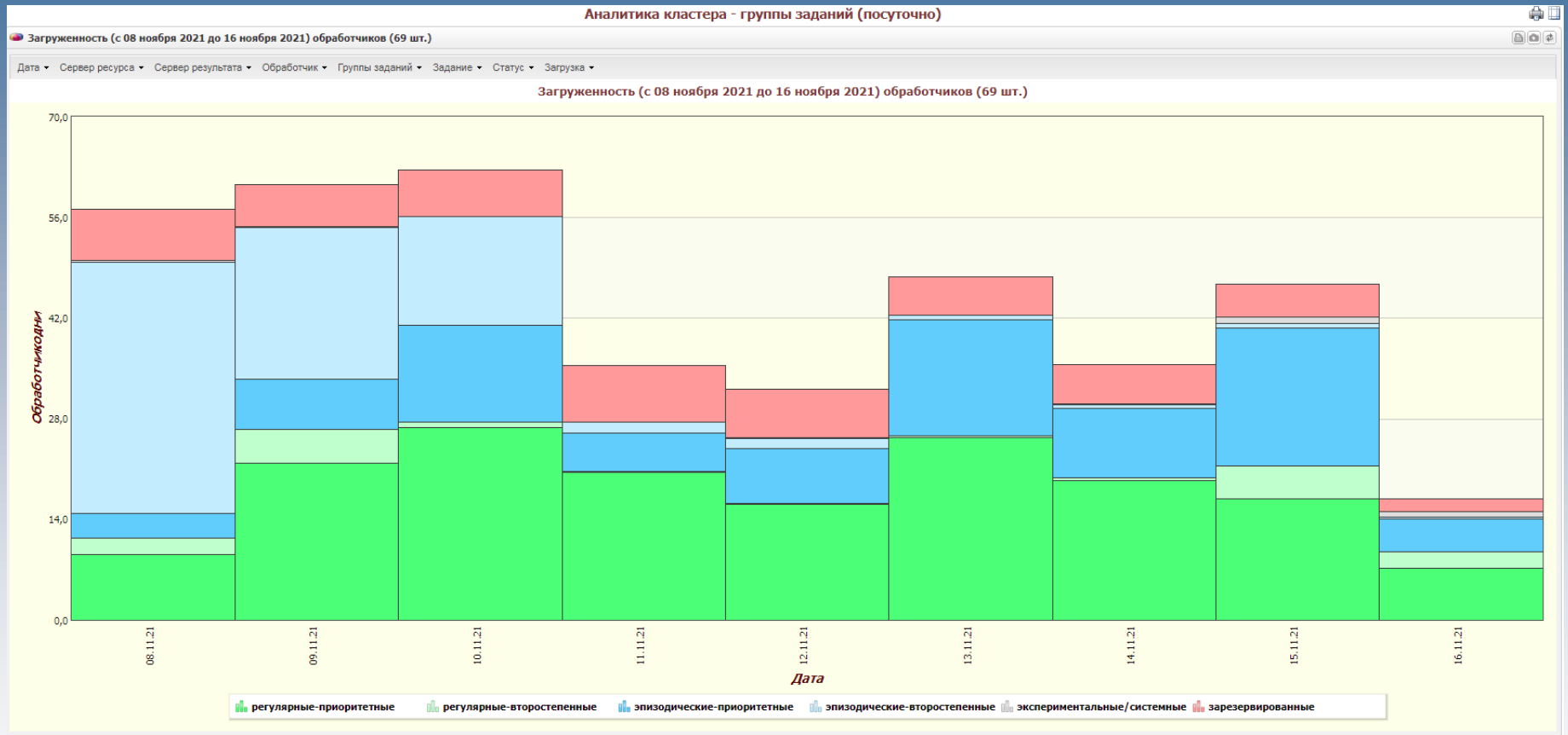
Настройка задач для обработчика

Обработчики имеющие это задание: PDL-6C-ENCELADUS

<< < >

Остальные обработчики: BLADE_06_GZ, BLADE_07_GZ, BLADE_09_GZ, BLADE_10_GZ, BLADE_11_GZ

Загруженность обработчиков по заданиям



Что влияет на скорость выполнения задачи

1.Скорость закачки данных на обработчик, увеличение скорости возможно если данные лежат на различных серверах – параллельная закачка данных

2.Скорость дискового массива обработчика – обычно именно эта характеристика выходит на первый план при любом виде распараллеливания частей обработки.

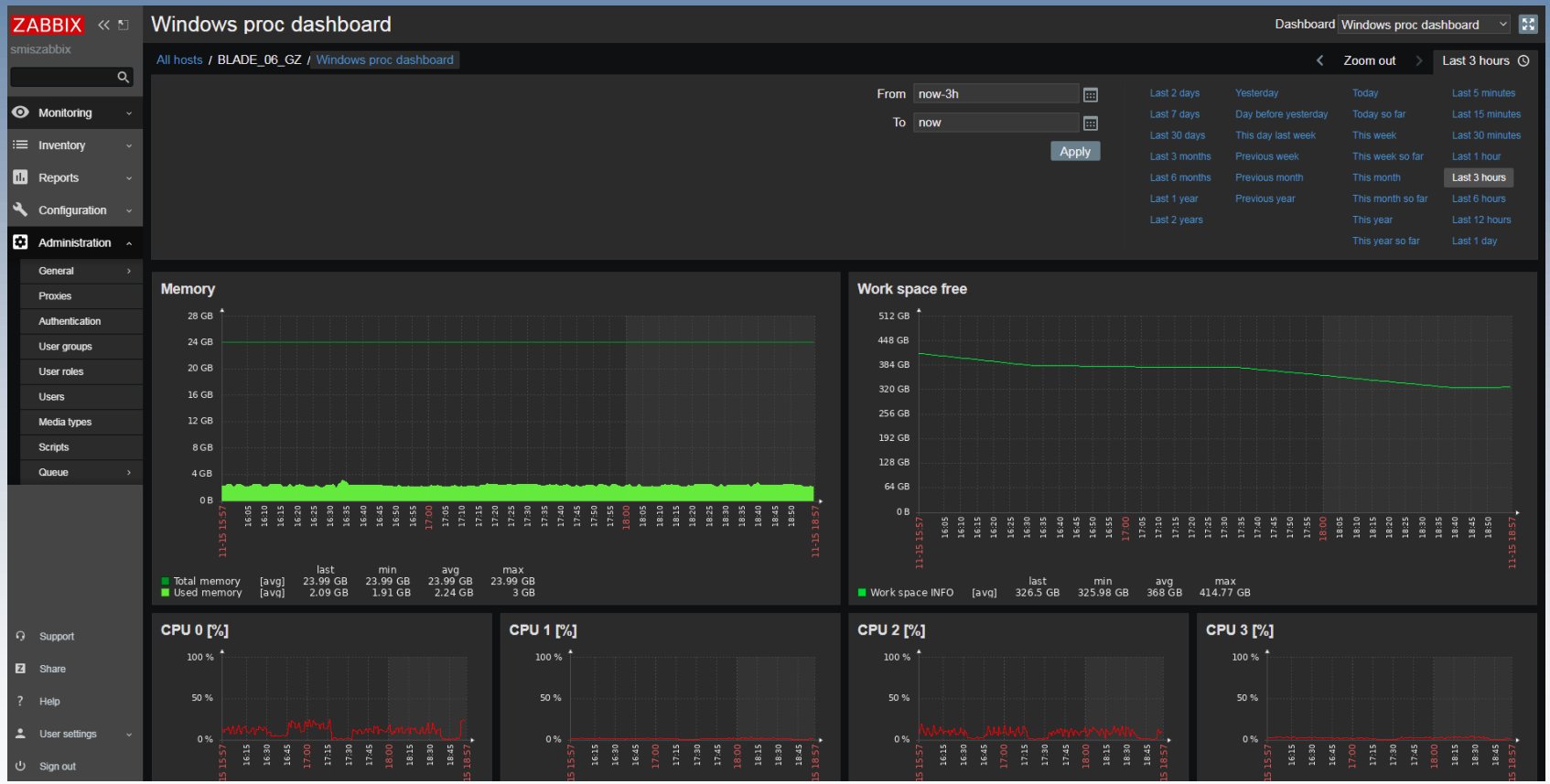
3. Скорость и количество оперативной памяти. Так как 32-разрядные приложения могут использовать только 3.5 гигабайта, можно распараллелить путем запуска нескольких приложений или нескольких заданий одновременно.

4.Скорость и количество ядер процессора. Основное место оптимизации для процессов, которые легко масштабируются потоками (threads), например каждый тред обрабатывает свою часть изображения.

5.Техническая грамотность кода процессов, участвующих в обработке.

Zabbix

Система позволяет контролировать различные параметры эксплуатации элементов аппаратно-программного комплекса.



https://habr.com/ru/post/485538

Эксперимент с виртуализацией

В рамках оптимизации использования мощностей обработки, проведем тесты по балансировке конкретных задач с использованием гипервизора VMWare ESXi версии 6.5
В качестве пилота выбран обработчик 2015 года покупки (REGOR).
Это U1 сервер Dell R220 (idrac 192.168.30.237) с характеристиками:

- Intel(R) Xeon(R) CPU E3-1241 v3 @ 3.50GHz (4 ядра, 8 потоков)
- 8G ОЗУ (один модуль DDR3 ECC 1600 MHz)
- Дисковый массив RAID1 из двух дисков 2Тб SATA 6Gb/s
- RAID-контроллер PERC H310
- Сетевая карта 1G

Что дает:
1. Снижает вероятность того, что физ. машина будет ничего не делать, если вылетит ошибка (на второй VM-ке будет работать)

2. Исключит ситуацию, когда обработчик кто-то берет под себя и потом на нем ничего не делает неделями (пока идет научный процесс придумывания/ваяния кода)

Таблица замеров:

Результаты тестирования

№	Показатель	**ndvi hitse16 geotif (session)** легкое	**swvi geotif (session)** легкое	**egorovs weekly slide mosaic terra recalc (session)** среднее	**create 4dc v05 terra (session)** среднее	**sentinel1_operative (session)** (у обработки ограничение на запуск максимум в один поток) тяжелое	ТТХ ресурсов
1	В среднем на физическом обработчике	9 потоков 0,4 мин/п	11 потоков 0,2 мин/п	2 потока 30,3 мин/п	4 потока 21,4 мин/п	1 поток 19,7 мин/п	8C физика 8R физика
2	Vmw-regor-h	2 потока 0,5 мин/п				1 поток 185 мин/по	6C (1000000S) 6R (1000S)
3	Vmw-regor-b	1 поток 0,9 мин/п				1 поток * Убита по времени	6C(1000S,2000RSV) 6R(500S)
4	Vmw-regor-h	1 поток 4 мин/п	1 поток 0,8 мин/п	1 поток ** 18,4 мин/п			6C (1000000S) 6R (1000S)
				1 поток 140 мин/п			
5	Vmw-regor-b			1 поток 21 мин/п	1 поток 19,7 мин/п		6C(1000S,2000RSV) 6R(1000S,1512RSV)
			3 потока 0,4 мин/п				
6							
7							

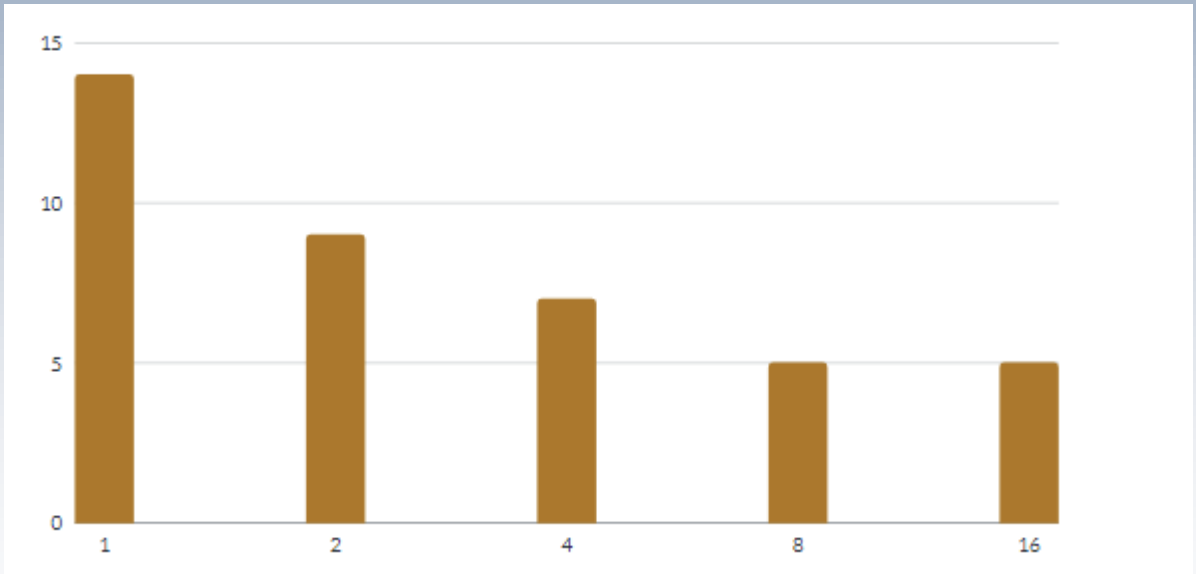
Результаты экспериментов

(*) sentinel1_operative (session) на Vmw-regor-b не может выполниться даже если Vmw-regor-h ничего не делает (превышается максимальный лимит времени на обработку). Если Vmw-regor-b не может подхватить обработку в случае когда Vmw-regor-h (не хватает располагаемых ресурсов), то после того как Vmw-regor-h освобождается, автоматически для Vmw-regor-b ресурсы не добавляются и обработку он все равно не подхватит. Вероятно ресурсы перераспределяются лишь в случае когда обработчик на них начинает претендовать, а для этого в его пуле должно быть легкое задание, которое запуститься даже в случае когда Vmw-regor-b обделен ресурсами.
(**) когда во второй конфигурации на Vmw-regor-h периодически (раз в 30 мин) запускалась легкая обработка, а на Vmw-regor-b постоянно средняя – все отрабатывало штатно, но попытка заменить легкую обработку на Vmw-regor-h на среднюю (оставив периодичность запусков) к успеху не привела, т.к. Vmw-regor-b постоянно делал среднюю обработку перетянув на себя ресурс от Vmw-regor-h, от чего на нем не осталось даже минимума доступного RAM, необходимого для запуска потока.

Выводы:
· Для работы такой схемы, нужно обязательно резервировать ресурсы на бэк (т.е. простой какой-то части ресурсов неизбежен в любом раскладе).
· Тяжелая обработка (красный тип), должна быть исключена из очереди для бэка
· 8Г ОЗУ, это самое минимальное кол-во, 16Г+ рекомендуется

Использование нескольких тредов в коде приложения обработки

Входные данные – 194 файла с данными MODIS (16.7 Gb).
Тип обработки – интерполяция
Выходные данные – 97 файлов, 8.7 Gb
Обработчик 16 ядер, 24 Gb оперативной памяти

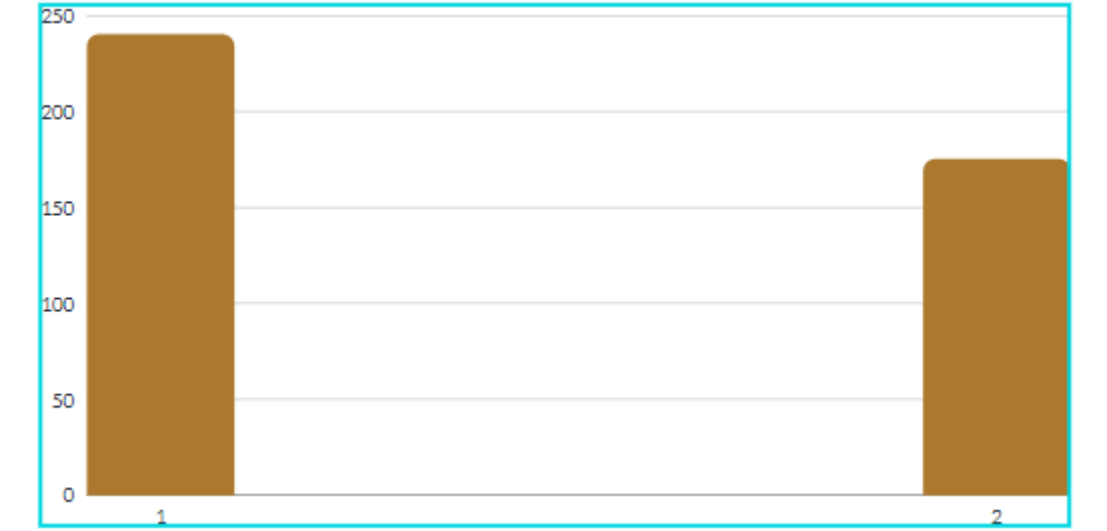


ИКИ

Разархивирование и перепаковка в бинарный формат

Входные данные – 2820 архивов (gz) с данными MODIS общим размером 111 gB.
Выходные данные – 8400 файлов, 235 gB

Скорость диска – 100 мб/сек
16-ядерный процессор

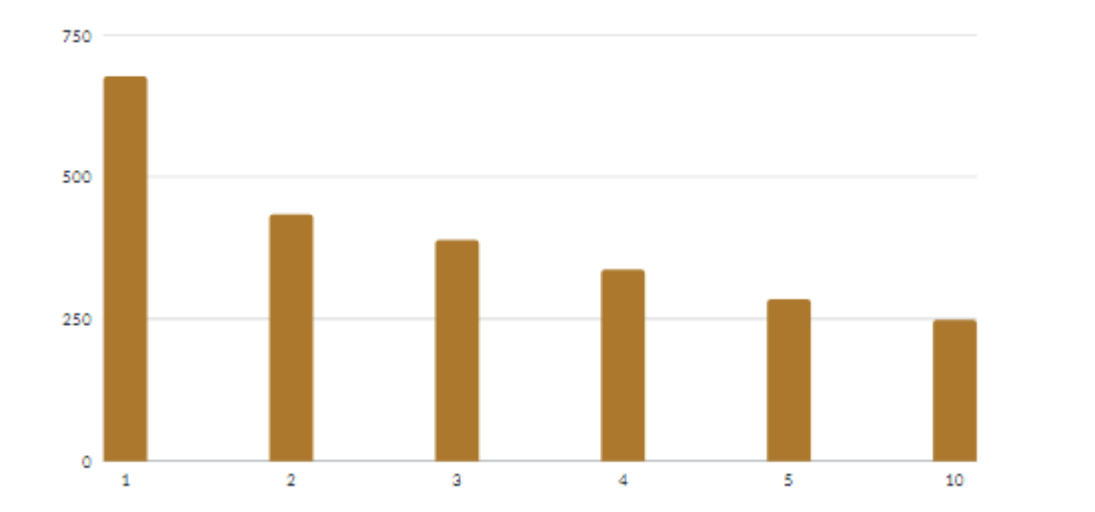


ИКИ

Параллельный запуск заданий

Входные данные – 20 заданий ~ 500 Mb
Выходные данные – 40 пригодных для отображения в картографическом интерфейсе тифов

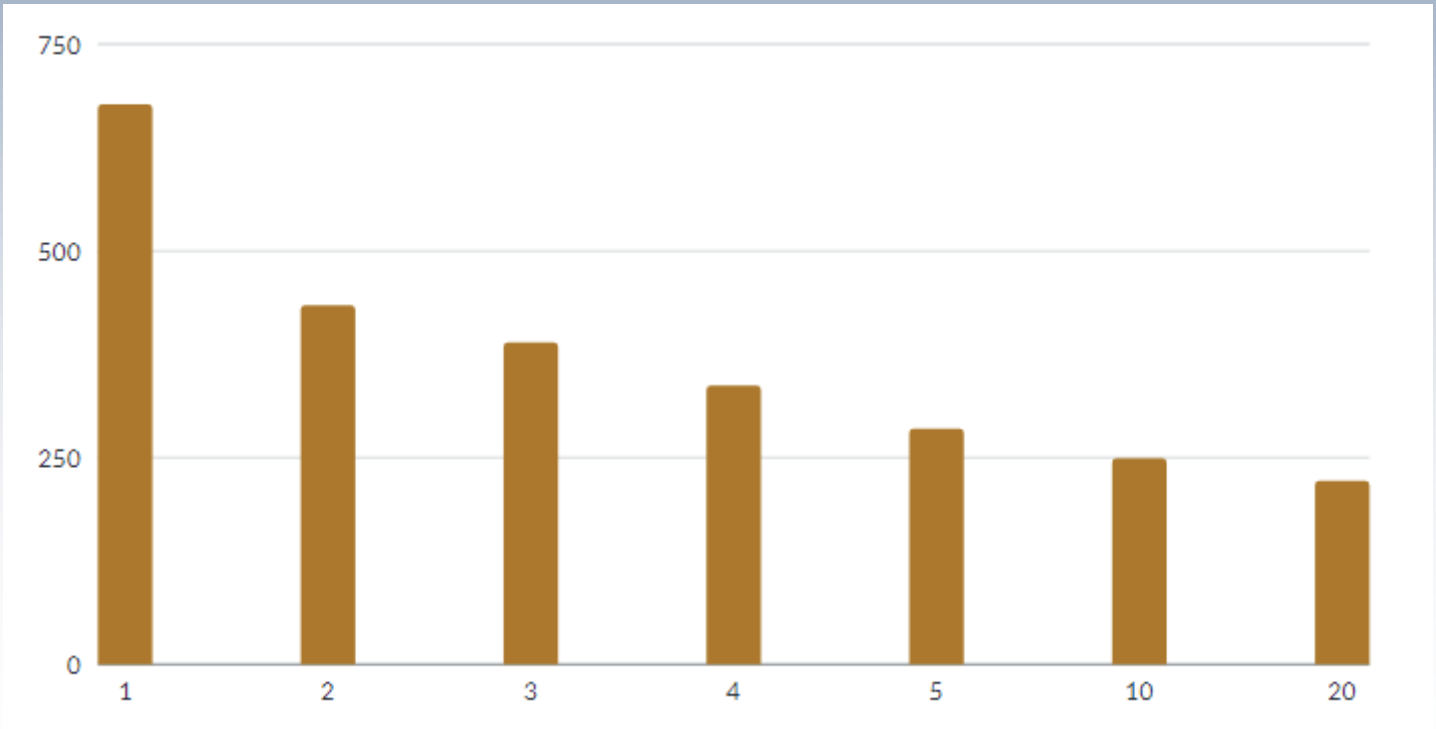
Скорость диска – 100 мб/сек
8-ядерный процессор
16 Gb оперативной памяти



ИКИ

Распараллеленное копирование

Хотя на маленьких масштабах (мало обработчиков, мало заданий) копирование в параллельном режиме дает некоторый выигрыш в скорости, при увеличении числа потоков, это приводит к проблемам отдачи данных на стороне сервера и скорость обработки начинает наоборот падать. На текущий момент копирование осуществляется последовательно для каждого задания и только потом параллелится.

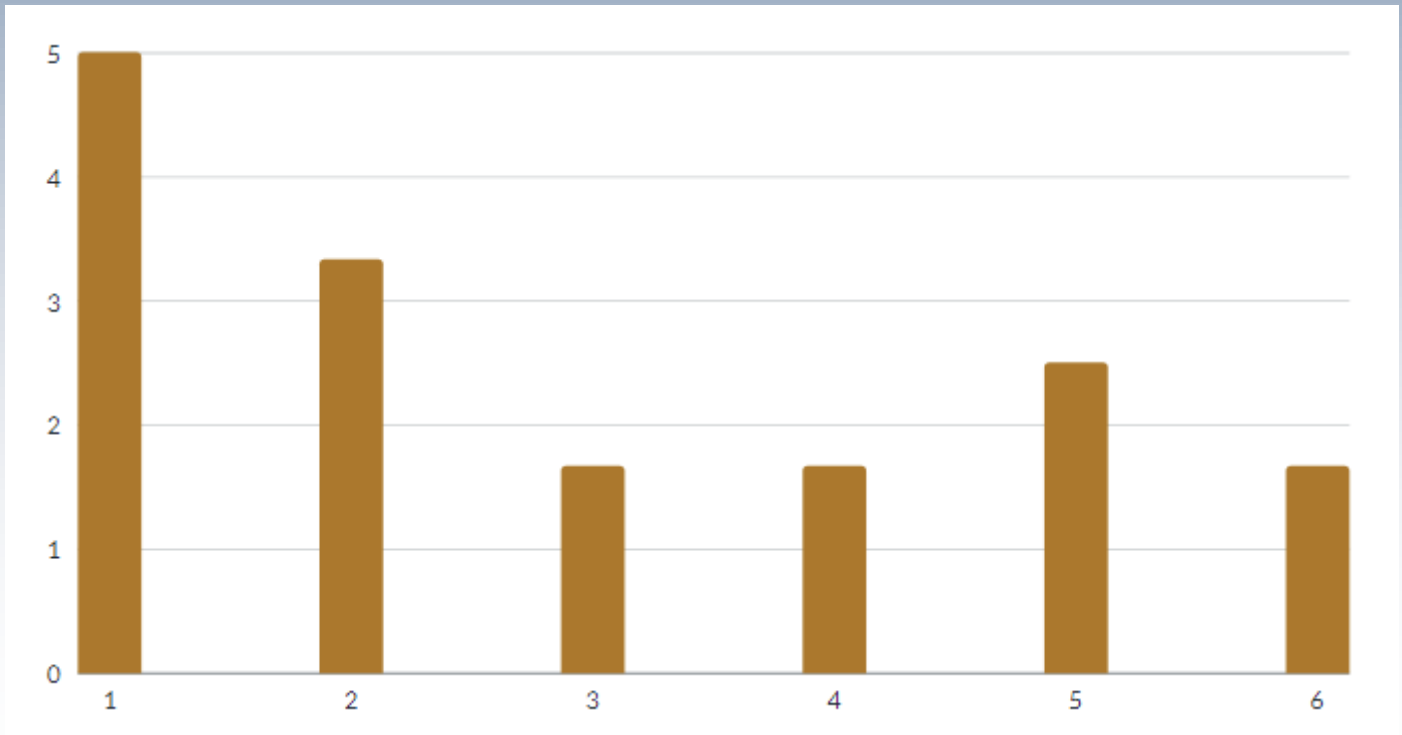


ИКИ

Избыточное распараллеливание

Входные данные – 400 сцен Landsat-7/8 ~ 150 Gb
Выходные данные – 400 файлов масок облачности, 680 Mb

Скорость диска – 100 мб/сек
12-ядерный процессор



BI+Zabbix

Интеграция этих систем позволяет получить мощный инструмент анализа состояния комплекса обработки. Ниже представлена загрузка обработчиков всеми выполняемыми заданиями.

Группы заданий	Задания	RAM (Гб)	CPU (ядра)	HDD (Гб)
⊞ регулярные-приоритетные		3,6	1,9	4,0
⊞ регулярные-второстепенные		1,0	1,4	3,5
⊞ эпизодические-приоритетные		0,4	1,0	1,6
⊞ эпизодические-второстепенные		1,1	1,1	1,2
⊟ экспериментальные/системные	⊞ benchmark hdf to geotif (session)	0,0	0,1	
	⊞ benchmark hitse interpolation modis (session)	7,0	5,0	227,6
	⊞ benchmark kmss hitse interpolation	1,6	3,3	10,4
⊟ зарезервированные	⊞ kmss optse interpolation (test)	1,0	4,9	9,1
	⊞ reserved			

ИКИ

Планы на будущее

1. Добавить отслеживание скорости отдачи дискового массива на обработчиках, получить таким образом полное понимание происходящего на обработчиках при полной загрузке заданиями.

2. Учесть полученные в ходе подготовки данного доклада результаты и оптимизировать работу ключевых и долгих процессов обработки.

3. Продолжить тестирование работы виртуалок на других конфигурациях реальных машин обработки

Спасибо за внимание

ИКИ