Исследование причин аварийного завершения заданий на суперкомпьютере “Ломоносов-2”

Соболев Сергей Игоревич

Лаборатория параллельных информационных технологий
НИВЦ МГУ имени М.В. Ломоносова

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ “СУПЕРКОМПЬЮТЕРНЫЕ ДНИ В РОССИИ”
Москва, МГУ, 26 сентября 2022 г.

О проекте

- Грант РФФИ “Разработка методов сохранения, реконструкции и анализа структурно-функциональных свойств суперкомпьютерных систем”
- Задача проекта – сохранение множества данных о состоянии компонентов суперкомпьютера, позволяющих оценить корректность их работы и влияние на работу суперкомпьютера в целом
- Основной объект – суперкомпьютер “Ломоносов-2”

Системные журналы “Ломоносова-2”

- Раньше сохранялись в файловой системе (узлы -> сервер)
- Налажено сохранение в InfluxDB (rsyslog+Telegraf)
- ~200 000 записей в час:
  - 81% – вычислительные узлы
  - 19% – служебные серверы

Модуль Log Viewer системы визуализации Chronograf

Простейший анализ

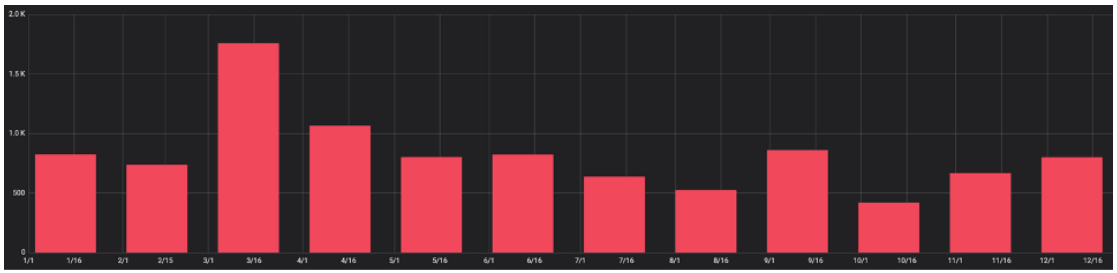
hostname	syslog.count_message
n49712	354 K
n52603	290
n52413	287
n52601	286
n52602	285
n52604	284
n50408	116
n53121	37
n53117	37
n51024	36
n49026	36
n53723	33
n50515	32
n53122	31
n52414	30
n50727	29
n52626	28
n50510	28

Узлы – записи за сутки

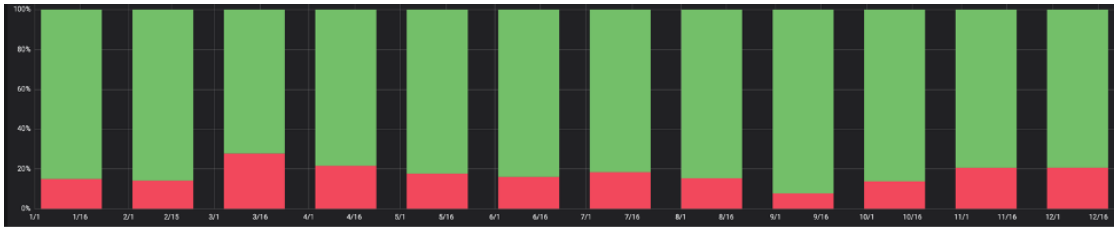
hostname	syslog.count_message
oss07	54 K
service-09	2 K
mon02	972
oss08	667
oss11	459
oss01	426
access-02	334
service-02	192
service-01	96
oss12	56
oss03	55
mds01	45
oss06	36
oss09	33
oss10	30
oss02	30
access-01	30
oss05	26
mds02	26

Серверы – записи за сутки

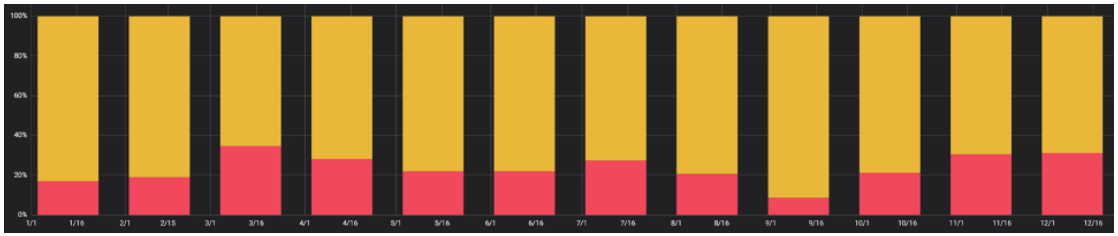
Поиск причин аварийного завершения заданий



Число задач со статусом FAILED по месяцам 2021 г.



Соотношение числа задач FAILED и задач с другими статусами



Соотношение числа задач FAILED и задач COMPLETED

Метод анализа

- Рассматриваются задачи со статусами FAILED, COMPLETED и NODE_FAIL
- Исключается раздел test
- Из системных журналов узлов выделяются события в районе завершения задачи:
  - для длинных задач – 15 минут до завершения и 3 минуты после завершения
  - для коротких задач – 3 минуты до начала и 3 минуты после завершения
- Список событий анализируется визуально
- Частые/важные события подсчитываются отдельно

Веб-интерфейс: выбор даты

Date	Tasks Processed *	FAILED Tasks	NODE_FAIL'ed Tasks	COMPLETED Tasks	Download Data
2021-11-12	67	Total: 30 warning 9 err 4 crit 2		Total: 37 warning 12 err 5 crit 3	csv
2021-11-13	55	Total: 18 warning 2 err 2 crit 2		Total: 37 warning 9 err 3 crit 1	csv
2021-11-14	64	Total: 15 warning 1 err 2		Total: 49 warning 12 err 11 crit 5	csv
2021-11-15	70	Total: 21 warning 2 err 1		Total: 49 warning 11 err 7	csv
2021-11-16	162	Total: 48 warning 8 err 16 crit 2	Total: 5 warning 1	Total: 109 warning 19 err 15 crit 4 alert 1	csv
2021-11-17	105	Total: 36 warning 8 err 2 crit 1		Total: 69 warning 9 err 10 crit 2	csv
2021-11-18	93	Total: 41 warning 4 err 9 crit 3		Total: 52 warning 7 err 5 crit 1	csv
2021-11-19	216	Total: 51 warning 11 err 2 crit 2		Total: 165 warning 19 err 100 crit 1 alert 1	csv
2021-11-20	80	Total: 12 warning 5 err 1 crit 1		Total: 68 warning 17 err 24	csv
2021-11-21	101	Total: 22 warning 6 err 6 crit 2		Total: 79 warning 37 err 22 crit 2	csv
2021-11-22	142	Total: 29 warning 6 err 2 crit 2 alert 1	Total: 2 warning 2 err 1 alert 1	Total: 111 warning 36 err 21 crit 1	csv
2021-11-23	267	Total: 44 warning 10 err 1 crit 3		Total: 223 warning 15 err 55 crit 8	csv
2021-11-24	123	Total: 16 warning 6 err 4	Total: 17 warning 3 err 1	Total: 90 warning 16 err 3 crit 5	csv

Веб-интерфейс: задачи за сутки

FAILED

Job Info									Node Log Entries								Node Typical Issues											Download
																	Software Faults								Hardware Faults				
job_id	t_start	t_end	run_time	node_list	nodes	command	partition	account	emerg	alert	crit	err	warning	notice	info	debug	segfault	orted segfault	dimmon segfault	xalt segfault	lustre	out of mem	blocked fs	PAM resolve symbol	mcelog	temperature	pcieport	NVRM	nodes logs
1339903	2021-11-11 09-45-08	2021-11-12 02-01-33	58585	n[49217-49218,49221-...	9	/opt/mpi/wrappers/ompi ./run_lmp_mpi-fin...	compute_prio	mikhailglagolev_2123						10	117	10			1										zip
1340109	2021-11-12 02-15-29	2021-11-12 02-21-33	364	n[48419-48421,48423,...	43	/home/mazalevaolya_2237/firefly/qdpt_dpb...	compute_prio	mazalevaolya_2237				120	867	43	303	50			49	4			60		18				zip
1340222	2021-11-12 05-34-28	2021-11-12 05-37-35	187	n[48109-48110,48605,...	15	/opt/mpi/wrappers/ompi start.sh	compute_prio	annglagoleva_2123						19	108	24			10										zip
1340204	2021-11-12 04-33-24	2021-11-12 05-38-17	3893	n[49714,49728,50220,...	15	/opt/mpi/wrappers/ompi /home/lazutin_212...	compute_prio	lazutin_2123					1	8	45	3			1										zip
1340051	2021-11-12 03-10-05	2021-11-12 06-57-28	13643	n[54325-54332,54427]	9	/opt/mpi/wrappers/ompi /mnt/scratch/user...	pascal	markinanasty_59					1	1	110	12			8	1									zip
1340253	2021-11-12 07-48-05	2021-11-12 07-53-26	321	n[48014,48017-48018,...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in re...	compute	osyzgantseva_2156					18	17	276	17	10		10										zip
1340254	2021-11-12 07-54-29	2021-11-12 07-57-57	208	n[48014,48017-48018,...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in re...	compute	osyzgantseva_2156					18	18	297	18	10		11										zip
1340012	2021-11-12 07-09-14	2021-11-12 08-38-00	5326	n[51619,52322]	2	/opt/mpi/wrappers/ompi ./batch4.sh	compute	bystrov_2221							2														zip
1340227	2021-11-12 10-21-32	2021-11-12 10-23-18	106	n[54119-54126,54422-...	12	/opt/mpi/wrappers/ompi vasp_gpu	pascal	vladimirs						24	139	74			23										zip
1340249	2021-11-12 10-23-44	2021-11-12 10-23-53	9	n[54201,54511-54512,...	5	/mnt/scratch/users/grigorenko/nwchem/ven...	pascal	grigorenko						8	38	16			4										zip
1340248	2021-11-12 10-23-44	2021-11-12 10-23-58	14	n[54119-54126,54422-...	12	/opt/mpi/wrappers/ompi vasp_gpu	pascal	vladimirs						24	140	74			23										zip
1340347	2021-11-12 13-03-34	2021-11-12 13-17-35	841	n[48030,48407-48408,...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in re...	compute	osyzgantseva_2156						1	40	2			1										zip
1340348	2021-11-12 13-04-51	2021-11-12 13-21-59	1028	n[49519-49520,49617-...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in re...	compute	osyzgantseva_2156							18														zip
1340364	2021-11-12 15-08-58	2021-11-12 15-23-45	887	n[48615,49104,49128,...	18	/opt/mpi/wrappers/ompi pw.x -nk 2 -in vc...	compute	osyzgantseva_2156						9	74	9			9										zip
1340366	2021-11-12 15-21-09	2021-11-12 15-30-14	545	n[49527,49614,49620,...	18	/opt/mpi/wrappers/ompi pw.x -nk 2 -in vc...	compute	osyzgantseva_2156						17	155	18			27										zip
1340273	2021-11-12 15-30-15	2021-11-12 15-31-14	59	n[48101,48615,49104,...	72	/opt/mpi/wrappers/ompi RBC -parallel	compute_prio	vasilieva_2242						74	597	78			106										zip
1340336	2021-11-12 15-31-20	2021-11-12 15-35-57	277	n[48101,49719-49722,...	15	/opt/mpi/wrappers/ompi /home/lazutin_212...	compute_prio	lazutin_2123						35	191	38			29										zip
1339948	2021-11-11 15-14-50	2021-11-12 15-38-00	87790	n[48214,48429,48714,...	14	/opt/mpi/wrappers/ompi RBC -parallel	compute_prio	rodion_2242					1	15	69	17			8	4									zip
1336834	2021-11-12 15-31-20	2021-11-12 15-38-09	409	n[49107-49108,49123-...	22	/mnt/scratch/users/kabylda_491/ff9/13_21...	compute	kabylda_491						20	199	20			26										zip
1340370	2021-11-12 15-38-26	2021-11-12 15-38-36	10	n[52120,52421]	2	ompi	compute	zodiacnv_1977																					zip
1340365	2021-11-12 15-36-31	2021-11-12 15-47-06	635	n[48101,49719-49722,...	18	/opt/mpi/wrappers/ompi pw.x -nk 2 -in vc...	compute	osyzgantseva_2156						11	87	10			5										zip
1340377	2021-11-12 16-00-19	2021-11-12 16-00-25	6	n[52120,52421]	2	ompi	compute	zodiacnv_1977						5	59	6			5										zip
1340380	2021-11-12 16-02-59	2021-11-12 16-03-03	4	n[52120,52421]	2	srun	compute	zodiacnv_1977			2	2		4	57	4			4										zip
1340383	2021-11-12 16-06-25	2021-11-12 16-06-31	6	n[52120,52421]	2	mpirun	compute	zodiacnv_1977			2	2		4	56	4			4										zip
1337880	2021-11-11 07-14-10	2021-11-12 16-08-14	118444	n[48005,48010,48013,...	14	/opt/mpi/wrappers/ompi RBC -parallel	compute	rodion_2242					1	15	331	15			13	4									zip
1340391	2021-11-12 17-00-09	2021-11-12 17-05-30	321	n[48129-48130,48317-...	8	/opt/mpi/wrappers/ompi /mnt/scratch/user...	compute	bugaevaas18_1247				40	300	5	226	6	10		6				20						zip
1340287	2021-11-12 11-58-45	2021-11-12 22-17-19	37114	n[48231,48503-48504,...	15	/opt/mpi/wrappers/ompi start.sh	compute_prio	annglagoleva_2123						5	27	5			1										zip
1340335	2021-11-12 23-11-21	2021-11-12 23-11-36	15	n[54220-54227,54232,...	24	/opt/mpi/wrappers/ompi /mnt/scratch/user...	pascal	rusmamont_59					3	23	329	52			33										zip
1340473	2021-11-12 23-45-56	2021-11-12 23-50-20	264	n[48513-48514,49525-...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in re...	compute	osyzgantseva_2156						12	72	12			12										zip
1340476	2021-11-12 23-54-45	2021-11-12 23-54-55	10	n[48513-48514,49525-...	18	/opt/mpi/wrappers/ompi pw.x -nk 4 -in sc...	compute	osyzgantseva_2156						17	285	17			17										zip

COMPLETED

Job Info									Node Log Entries								Node Typical Issues											Download
																	Software Faults								Hardware Faults				
job_id	t_start	t_end	run_time	node_list	nodes	command	partition	account	emerg	alert	crit	err	warning	notice	info	debug	segfault	orted segfault	dimmon segfault	xalt segfault	lustre	out of mem	blocked fs	PAM resolve symbol	mcelog	temperature	pcieport	NVRM	nodes logs
1339800	2021-11-11 07-09-07	2021-11-12 00-17-01	61674	n[54515-54518]	4	/home/vytas_2091/_scratch/opt/l2-multimd...	pascal	vytas_2091					2	5	57	10			3										zip
1340108	2021-11-11 21-18-40	2021-11-12 02-03-30	17090	n[48419-48421,48423,...	33	/mnt/scratch/users/sudarkovasveta_2237/f...	compute_prio	sudarkovasveta_2237					31	9	664	9			1	2	47								zip
1340075	2021-11-11 21-17-01	2021-11-12 02-11-05	17644	n54427	1	/opt/mpi/wrappers/ompi /mnt/scratch/user...	pascal	rusmamont_59						1	12	2													zip
1340239	2021-11-11 22-16-05	2021-11-12 02-15-28	14363	n[51504,51619,52210,...	6	/opt/mpi/wrappers/impi ./p_l hcch-nh3.cc...	compute	bulgacov2012_2215					3	5	83	5			4	4									zip
1339947	2021-11-11 13-36-01	2021-11-12 02-47-26	47485	n[54310-54313,54525-...	12	/opt/mpi/wrappers/ompi vasp_gpu	pascal	vladimirs						14	54	28			3										zip
1340212	2021-11-11 22-08-27	2021-11-12 03-13-37	18310	n50523	1	./run_gromacs_2.sh	compute	vvvas_2123							12														zip
1340252	2021-11-11 23-11-46	2021-11-12 03-38-09	15983	n[48321,49314,49322,...	18	/opt/mpi/wrappers/ompi pw.x -nk 2 -in sc...	compute	osyzgantseva_2156						11	50	11			10										zip
1339888	2021-11-11 07-08-55	2021-11-12 04-33-18	77063	n[48207,48219,48430-...	28	/mnt/scratch/users/eltsovia_2231/calcula...	compute_prio	eltsovia_2231						32	128	33			12										zip
1339929	2021-11-11 13-31-00	2021-11-12 05-34-23	57803	n[48109-48110,48605,...	16	/opt/mpi/wrappers/ompi vasp_std	compute_prio	klavsjuk_2016						17	67	17			1										zip
1339933	2021-11-11 14-52-41	2021-11-12 06-31-39	56338	n[48402-48403,48413-...	16	/opt/mpi/wrappers/ompi vasp_std	compute_prio	klavsjuk_2016					1	16	63	17			1										zip
1339900	2021-11-11 19-09-11	2021-11-12 07-06-15	43024	n52017	1	/opt/mpi/wrappers/run ./SLkMC111	compute	sergy2710_2016					3	1	24	1				3									zip
1339901	2021-11-11 19-11-35	2021-11-12 07-08-09	42994	n52322	1	/opt/mpi/wrappers/run ./SLkMC111	compute	sergy2710_2016					3	2	28	2			1	3									zip
1340038	2021-11-11 16-18-58	2021-11-12 08-49-01	59403	n54213	1	/opt/mpi/wrappers/ompi /mnt/scratch/user...	pascal	rusmamont_59						1	5	2			1										zip
1340281	2021-11-12 09-31-41	2021-11-12 09-54-22	1361	n53328	1	/mnt/scratch/users/mikhailnickkh_1704/FF...	compute	mikhailnickkh_1704						1	12	1													zip

Фиксируемые события

- Аварийное завершение процессов пользовательского приложения (segfault)
- Аварийное завершение системных процессов orted, dimmon, xalt_submis
- Проблемы с общей файловой системой Lustre
- Превышение приложением лимита доступной оперативной памяти (out of memory)
- Аппаратные проблемы, определяемые демоном mcelog
- Превышение заданного температурного порога ядром или конструктивом центрального процессора
- Ошибки системной шины PCIe узла
- Проблемы с графическим ускорителем NVIDIA

Первые результаты

- Большинство перечисленных проблем не являются критическими
  - … ни события с высоким уровнем важности (alert, critical)
  - … ни даже “segmentation fault” (ORCA, NWCHEM, VASP)
- Точную причину сбоя таким методом можно определить лишь для единичных заданий
- 100% ответа на вопрос нет :(

Интересное наблюдение

- 37 случаев аварийного завершения процесса orted (OpenMPI)
  - из них 29 - при работе GROMACS
    - но данные мы видим только в конце работы приложения…

Планы: расширение источников данных

- События из системных журналов части служебных серверов, необходимых для поддержания работы суперкомпьютера (oss и mds)
- События из буфера ядра операционной системы (dmesg)
- Счетчики сетевых ошибок

Спасибо за внимание!