BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Feb 18, 2014

A Diagnostic Utility For Analyzing Periods Of Degraded Job Performance

Joshi Fullop & Robert Sisneros

National Center for Supercomputing Applications

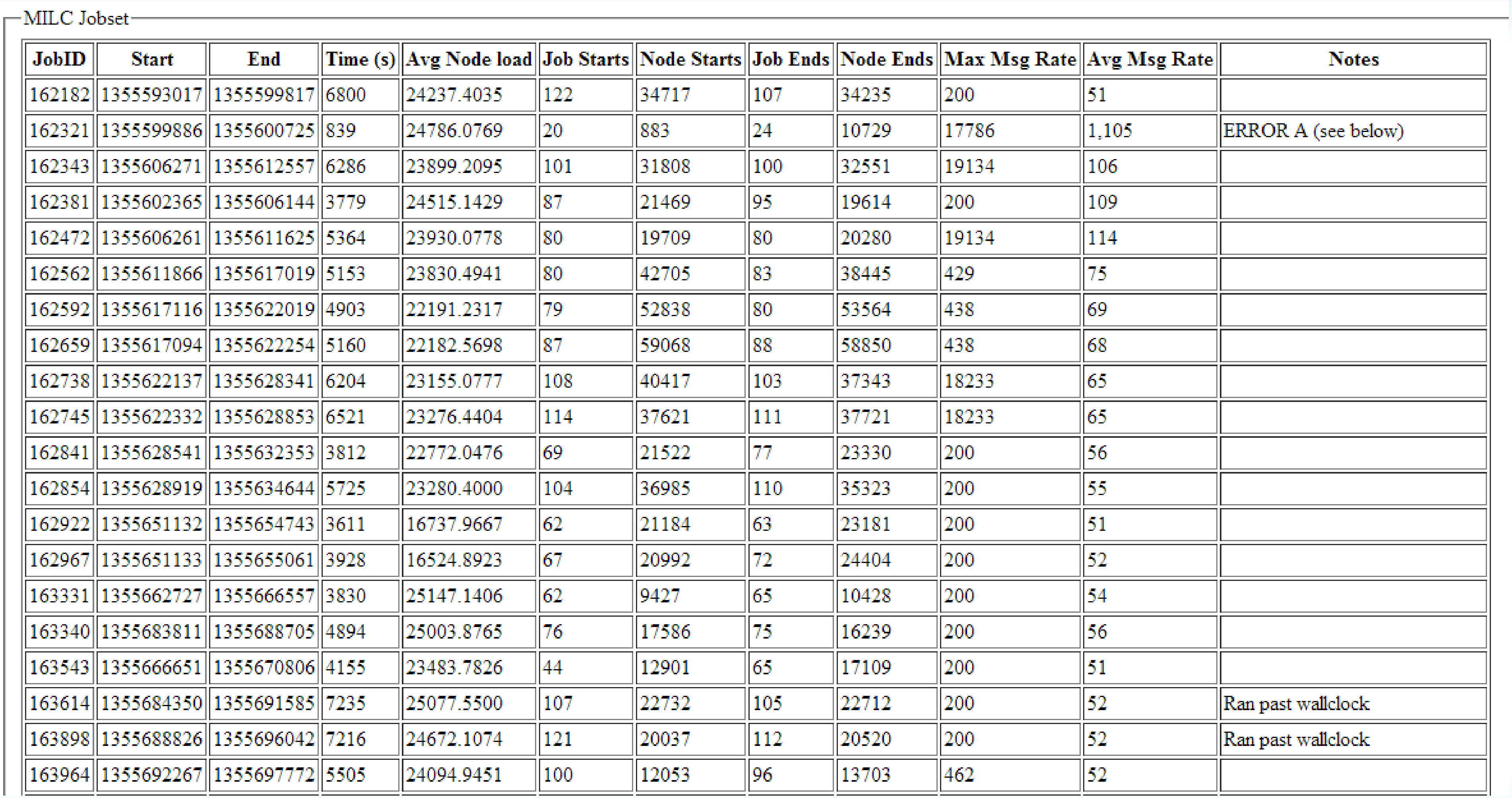
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Background

- We set out to answer a common question that is often posed in different ways depending on perspective.
- **User**: “Why did one of my jobs get better performance than another?”
- **Admin**: “Why did we get better throughput during one period than another?”
- **Researcher**: “How can we characterize the differences between two periods?”

Background

- Based on a consistency analysis study group comprised of members of Cray and NCSA.
- Ran multiple codes with standardized inputs many times over a study period.
- Charted application run times vs many different measurable variables.
- First step: Job stats and direct metric analysis.



MILC Jobset

JobID	Start	End	Time (s)	Avg Node load	Job Starts	Node Starts	Job Ends	Node Ends	Max Msg Rate	Avg Msg Rate	Notes
162182	1355593017	1355599817	6800	24237.4035	122	34717	107	34235	200	51	
162321	1355599886	1355600725	839	24786.0769	20	883	24	10729	17786	1,105	ERROR A (see below)
162343	1355606271	1355612557	6286	23899.2095	101	31808	100	32551	19134	106	
162381	1355602365	1355606144	3779	24515.1429	87	21469	95	19614	200	109	
162472	1355606261	1355611625	5364	23930.0778	80	19709	80	20280	19134	114	
162562	1355611866	1355617019	5153	23830.4941	80	42705	83	38445	429	75	
162592	1355617116	1355622019	4903	22191.2317	79	52838	80	53564	438	69	
162659	1355617094	1355622254	5160	22182.5698	87	59068	88	58850	438	68	
162738	1355622137	1355628341	6204	23155.0777	108	40417	103	37343	18233	65	
162745	1355622332	1355628853	6521	23276.4404	114	37621	111	37721	18233	65	
162841	1355628541	1355632353	3812	22772.0476	69	21522	77	23330	200	56	
162854	1355628919	1355634644	5725	23280.4000	104	36985	110	35323	200	55	
162922	1355651132	1355654743	3611	16737.9667	62	21184	63	23181	200	51	
162967	1355651133	1355655061	3928	16524.8923	67	20992	72	24404	200	52	
163331	1355662727	1355666557	3830	25147.1406	62	9427	65	10428	200	54	
163340	1355683811	1355688705	4894	25003.8765	76	17586	75	16239	200	56	
163543	1355666651	1355670806	4155	23483.7826	44	12901	65	17109	200	51	
163614	1355684350	1355691585	7235	25077.5500	107	22732	105	22712	200	52	Ran past wallclock
163898	1355688826	1355696042	7216	24672.1074	121	20037	112	20520	200	52	Ran past wallclock
163964	1355692267	1355697772	5505	24094.9451	100	12053	96	13703	462	52	

Direct Correlation Analysis

- Findings:
  - There were no ‘very strong’ direct correlations with any single variable.
  - With as complex an environment like Blue Waters, this is not overly surprising.
  - Too many moving parts that too often depend on or are affected directly or indirectly by each other.

Next step: Log Analysis

- What do we have to work with?
  - Job records (Torque/Moab logs)
  - Systems logs (LLM)
  - Systems logs (ESMS)
  - Systems logs (Sonexions)
  - Systems logs (HPSS)
  - Systems logs (networking)
  - Systems logs (et alius)
  - NOT performance counters (yet)

Enabling Technology

- Hierarchical Event Log Organizer (HELO)
  - Machine learning system that classifies log messages and dynamically identifies new ones.
  - Tags each identified message with Template ID.
  - Manages Templates, automatically modifies them to include new, but similar log messages.
  - Summarized event count metadata is used to quickly compare log messages from different periods.

HELO enhancements

- Dynamic reordering of Template data structure in HELO online handler.
  - More frequently encountered templates get moved to the front of the list and are therefore found more quickly.
  - Quick response to surprise event storms.
- Dynamic Template Deactivation
  - Templates with no actual occurrences in a period of time get dropped from the active list for consideration, but not deleted.
  - They will be found in the server-side process when the online processor fails to find it.

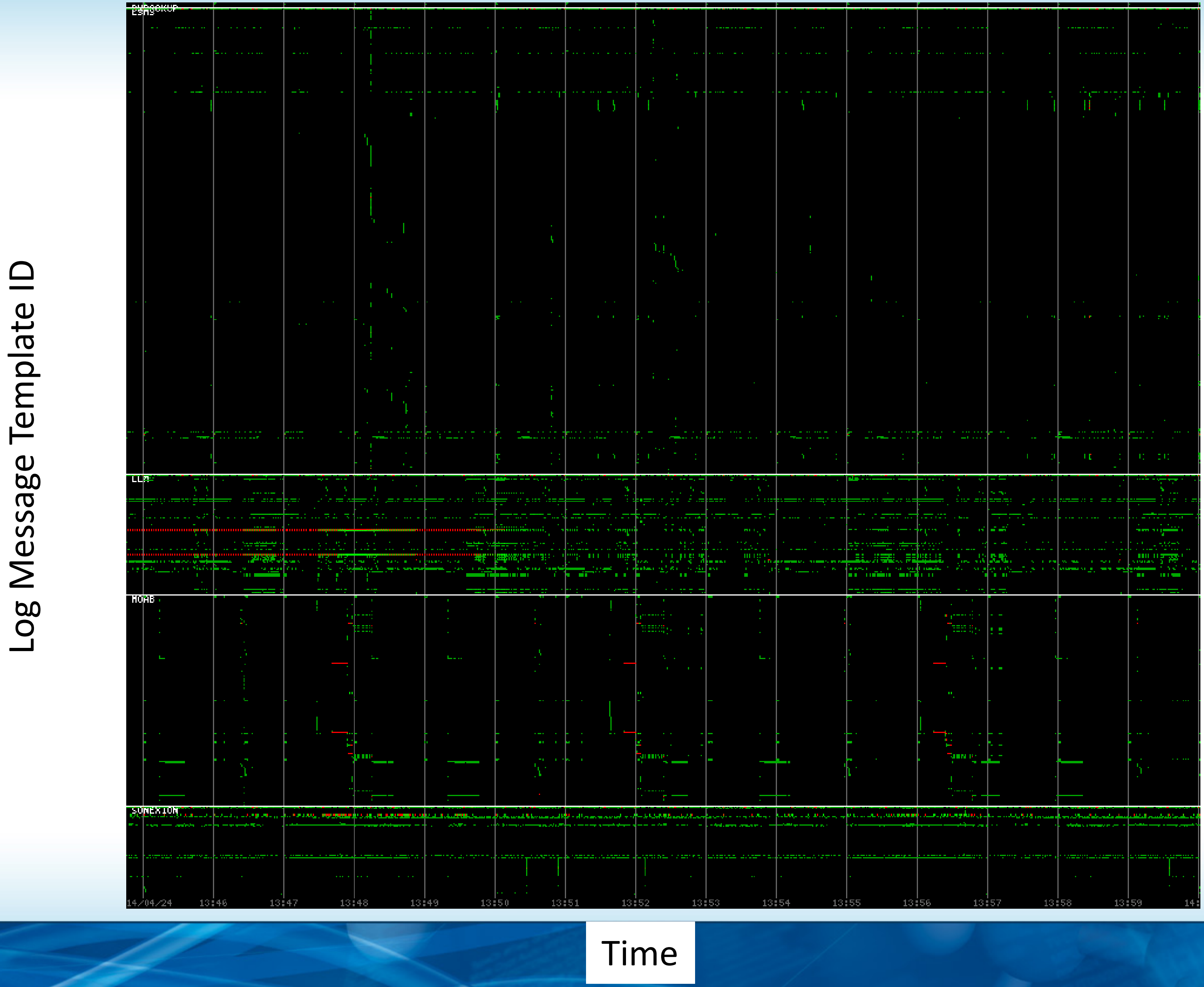
Premise

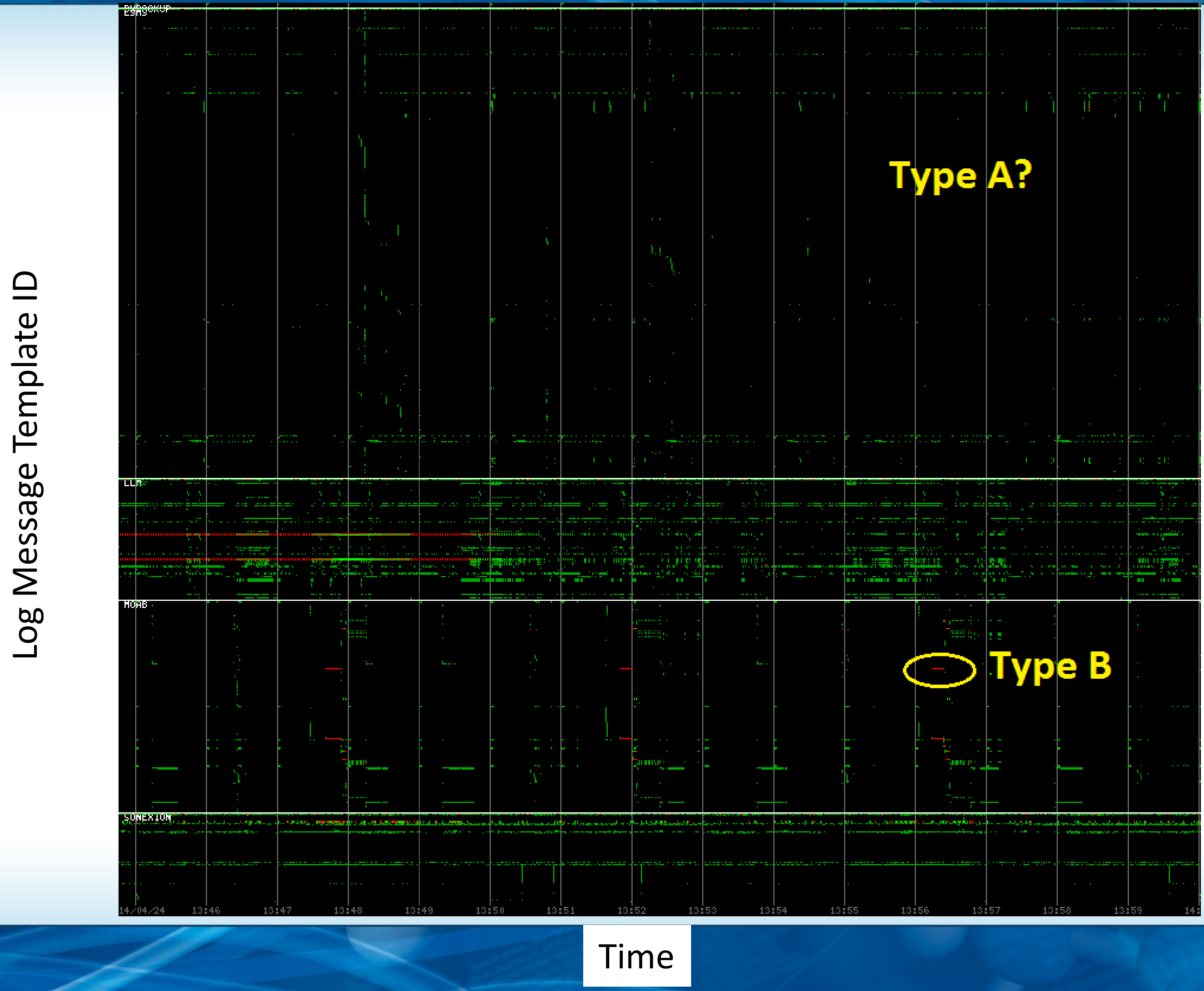
- Log messages are generated on a regular cycle (statistic reporting) or when a problem arises.
- A stressed system will generate more log traffic than healthy one.

- However, in the extreme opposite case, i.e. when there is NO log traffic… things are gravely wrong.

Log Message Patterns

- Single Event Per Failure (Type A)
  - This is what the rest of the non-log processing world thinks exists to indicate a single point of failure, but so rarely do.
  - When they do occur, we certainly want to know.
  - Simple comparison: Occurred in one period but not the other.
- Multiple Similar Events Per Failure (Type B)
  - In systems with many multiples of the same reporting component, it is common for many if not all components to report the same problem.
  - Comparison: Ratio of occurrence count between periods


Log Message Template ID
DUPGOKUP
ESMS
LLM
MOAB
SONEXION
14/04/24
13:46
13:47
13:48
13:49
13:50
13:51
13:52
13:53
13:54
13:55
13:56
13:57
13:58
13:59
14:
Time


Log Message Template ID
ESMS
LLM
MOAB
SONEXION
14/04/24
13:46
13:47
13:48
13:49
13:50
13:51
13:52
13:53
13:54
13:55
13:56
13:57
13:58
13:59
14:
Time

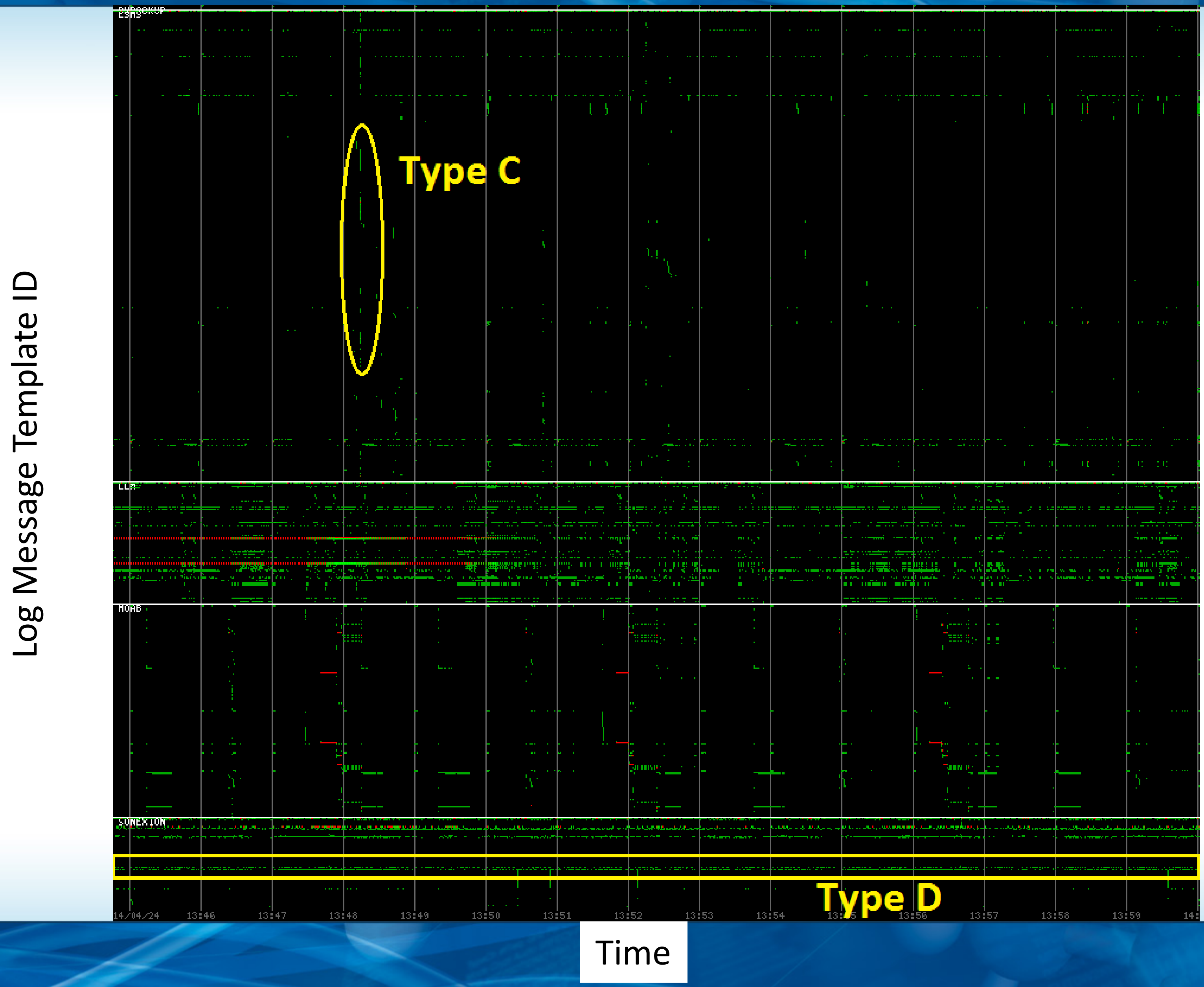
Type A?

Type B

Log Message Patterns

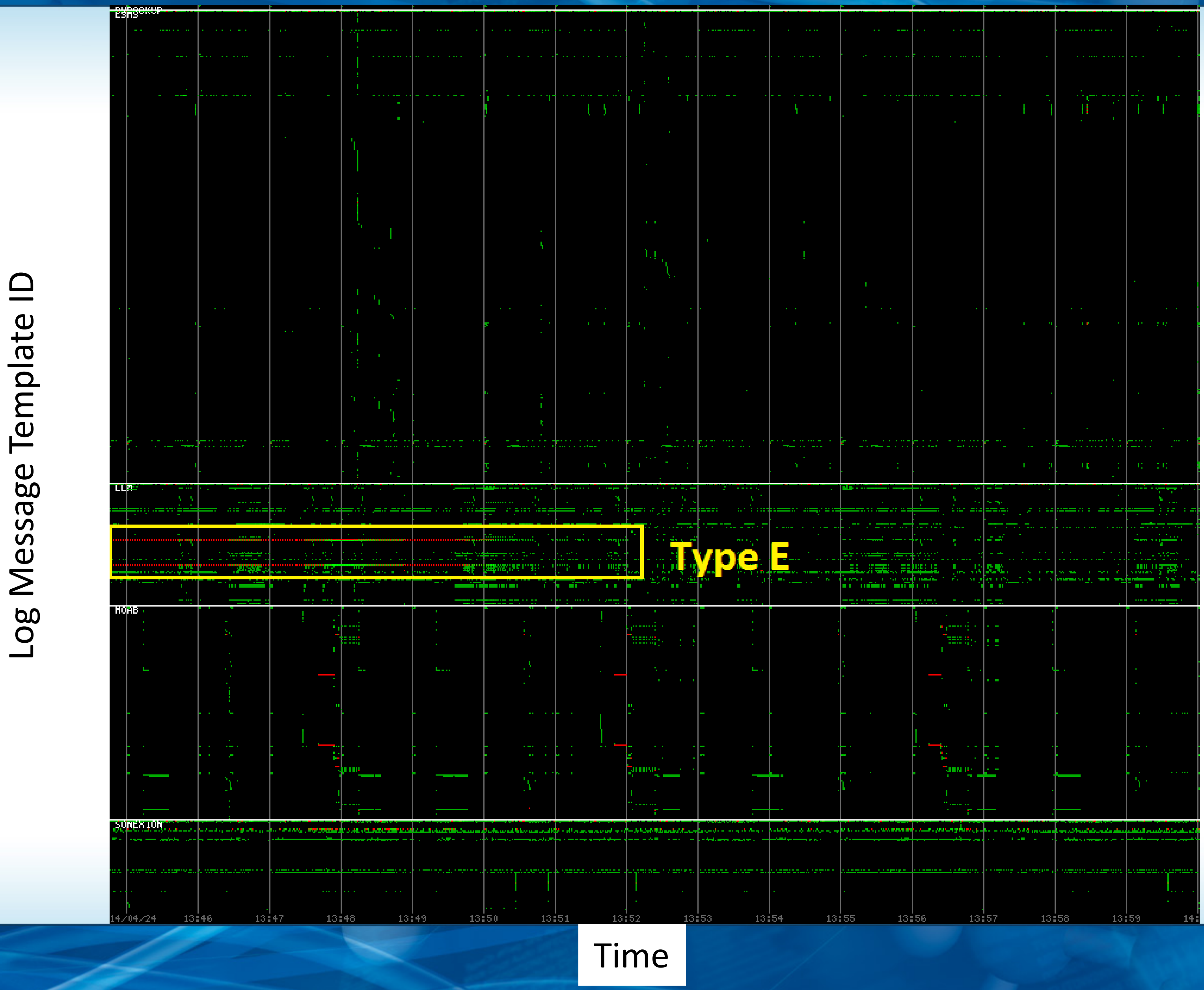
- Multiple Different Events Per Failure (Type C)
  - Often, when a failure occurs, many different log message appear. Some refer to this as an underlying failure’s fingerprint.
  - Our method considers a quasi-fingerprint as we do not attempt to consider timing, just occurrence counts.
  - Comparison: Ratio of occurrence count between periods.

- Constant Rate Events (Type D)
  - Quintessential example: cron jobs
  - Comparison: ratio of event occurrence rates (counts normalized by period length)


Log Message Template ID
ESMS
Type C
LLM
MOAB
SONEXION
Type D
14/04/24
13:46
13:47
13:48
13:49
13:50
13:51
13:52
13:53
13:54
13:56
13:57
13:58
13:59
14:
Time

Log Message Patterns

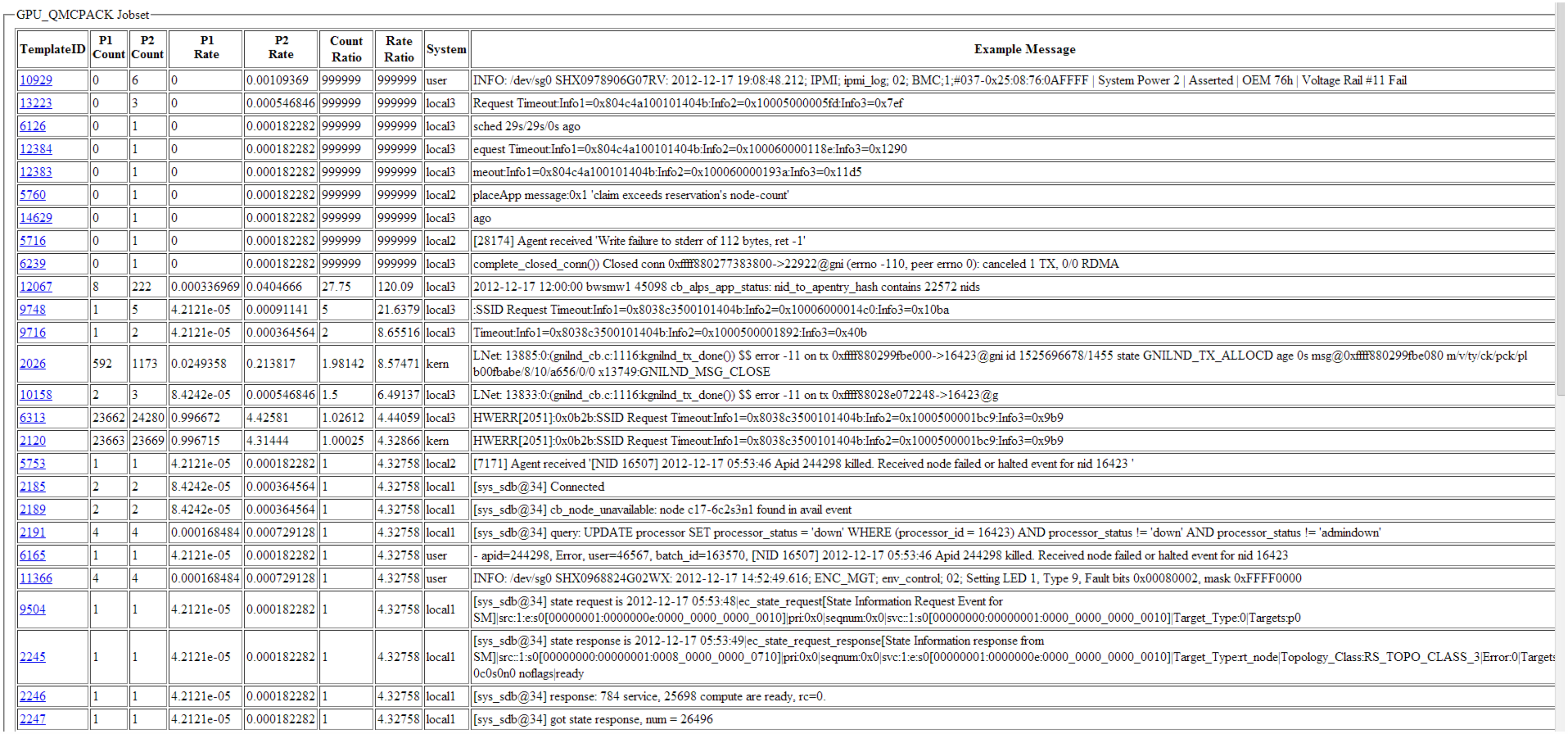
- Variable Rate Events (Type E)
  - Very difficult to notice variation with cursory human log analysis. The message will be seen in both periods, and usually discounted.
  - A certain log rate my be considered expected when components are operating in a normal mode, but heightened rates could indicate system distress.
  - Comparison: Ratio of event occurrence rates.


Log Message Template ID
ESMS
LLM
MOAB
SONEXION
Type E
14/04/24
13:46
13:47
13:48
13:49
13:50
13:51
13:52
13:53
13:54
13:55
13:56
13:57
13:58
13:59
14:
Time

Processing

1. Find set of events that occurred in the first period and summarize.
2. Complement with set of events that occurred in the second period and summarize.
3. Normalize counts to event rates for each period.
4. Calculate ratios of occurrence and rate ratios.
5. Sort by rate ratio, then occurrence count.

Output



GPU_QMCPACK Jobset

TemplateID	P1 Count	P2 Count	P1 Rate	P2 Rate	Count Ratio	Rate Ratio	System	Example Message
10929	0	6	0	0.00109369	999999	999999	user	INFO: /dev/sg0 SHX0978906G07RV: 2012-12-17 19:08:48.212; IPMI; ipmi_log; 02; BMC;1;#037-0x25:08:76:0AFFFF | System Power 2 | Asserted | OEM 76h | Voltage Rail #11 Fail
13223	0	3	0	0.000546846	999999	999999	local3	Request Timeout:Info1=0x804c4a100101404b:Info2=0x10005000005fd:Info3=0x7ef
6126	0	1	0	0.000182282	999999	999999	local3	sched 29s/29s/0s ago
12384	0	1	0	0.000182282	999999	999999	local3	equest Timeout:Info1=0x804c4a100101404b:Info2=0x10006000118e:Info3=0x1290
12383	0	1	0	0.000182282	999999	999999	local3	meout:Info1=0x804c4a100101404b:Info2=0x100060000193a:Info3=0x11d5
5760	0	1	0	0.000182282	999999	999999	local2	placeApp message:0x1 'claim exceeds reservation's node-count'
14629	0	1	0	0.000182282	999999	999999	local3	ago
5716	0	1	0	0.000182282	999999	999999	local2	[28174] Agent received 'Write failure to stderr of 112 bytes, ret -1'
6239	0	1	0	0.000182282	999999	999999	local3	complete_closed_conn()) Closed conn 0xffff880277383800->22922@gni (errno -110, peer errno 0): canceled 1 TX, 0/0 RDMA
12067	8	222	0.000336969	0.0404666	27.75	120.09	local3	2012-12-17 12:00:00 bwsmw1 45098 cb_alps_app_status: nid_to_apentry_hash contains 22572 nids
9748	1	5	4.2121e-05	0.00091141	5	21.6379	local3	:SSID Request Timeout:Info1=0x8038c3500101404b:Info2=0x10006000014c0:Info3=0x10ba
9716	1	2	4.2121e-05	0.000364564	2	8.65516	local3	Timeout:Info1=0x8038c3500101404b:Info2=0x100050001892:Info3=0x40b
2026	592	1173	0.0249358	0.213817	1.98142	8.57471	kern	LNet: 13885:0:(gnilnd_cb.c:1116:kgnilnd_tx_done()) $$ error -11 on tx 0xffff880299fbe000->16423@gni id 1525696678/1455 state GNILND_TX_ALLOCD age 0s msg@0xffff880299fbe080 m/v/ty/ck/pck/pl b00fbabe/8/10/a656/0/0 x13749:GNILND_MSG_CLOSE
10158	2	3	8.4242e-05	0.000546846	1.5	6.49137	local3	LNet: 13833:0:(gnilnd_cb.c:1116:kgnilnd_tx_done()) $$ error -11 on tx 0xffff88028e072248->16423@g
6313	23662	24280	0.996672	4.42581	1.02612	4.44059	local3	HWERR[2051]:0x0b2b:SSID Request Timeout:Info1=0x8038c3500101404b:Info2=0x1000500001bc9:Info3=0x9b9
2120	23663	23669	0.996715	4.31444	1.00025	4.32866	kern	HWERR[2051]:0x0b2b:SSID Request Timeout:Info1=0x8038c3500101404b:Info2=0x1000500001bc9:Info3=0x9b9
5753	1	1	4.2121e-05	0.000182282	1	4.32758	local2	[7171] Agent received '[NID 16507] 2012-12-17 05:53:46 Apid 244298 killed. Received node failed or halted event for nid 16423 '
2185	2	2	8.4242e-05	0.000364564	1	4.32758	local1	[sys_sdb@34] Connected
2189	2	2	8.4242e-05	0.000364564	1	4.32758	local1	[sys_sdb@34] cb_node_unavailable: node c17-6c2s3n1 found in avail event
2191	4	4	0.000168484	0.000729128	1	4.32758	local1	[sys_sdb@34] query: UPDATE processor SET processor_status = 'down' WHERE (processor_id = 16423) AND processor_status != 'down' AND processor_status != 'admindown'
6165	1	1	4.2121e-05	0.000182282	1	4.32758	user	- apid=244298, Error, user=46567, batch_id=163570, [NID 16507] 2012-12-17 05:53:46 Apid 244298 killed. Received node failed or halted event for nid 16423
11366	4	4	0.000168484	0.000729128	1	4.32758	user	INFO: /dev/sg0 SHX0968824G02WX: 2012-12-17 14:52:49.616; ENC_MGT; env_control; 02; Setting LED 1, Type 9, Fault bits 0x00080002, mask 0xFFFF0000
9504	1	1	4.2121e-05	0.000182282	1	4.32758	local1	[sys_sdb@34] state request is 2012-12-17 05:53:48|ec_state_request[State Information Request Event for SM]|src:1:e:s0[00000001:0000000e:0000_0000_0000_0010]|pri:0x0|seqnum:0x0|svc::1:s0[00000000:00000001:0000_0000_0000_0010]|Target_Type:0|Targets:p0
2245	1	1	4.2121e-05	0.000182282	1	4.32758	local1	[sys_sdb@34] state response is 2012-12-17 05:53:49|ec_state_request_response[State Information response from SM]|src::1:s0[00000000:00000001:0008_0000_0000_0710]|pri:0x0|seqnum:0x0|svc:1:e:s0[00000001:0000000e:0000_0000_0000_0010]|Target_Type:rt_node|Topology_Class:RS_TOPO_CLASS_3|Error:0|Targets 0c0s0n0 noflags|ready
2246	1	1	4.2121e-05	0.000182282	1	4.32758	local1	[sys_sdb@34] response: 784 service, 25698 compute are ready, rc=0.
2247	1	1	4.2121e-05	0.000182282	1	4.32758	local1	[sys_sdb@34] got state response, num = 26496

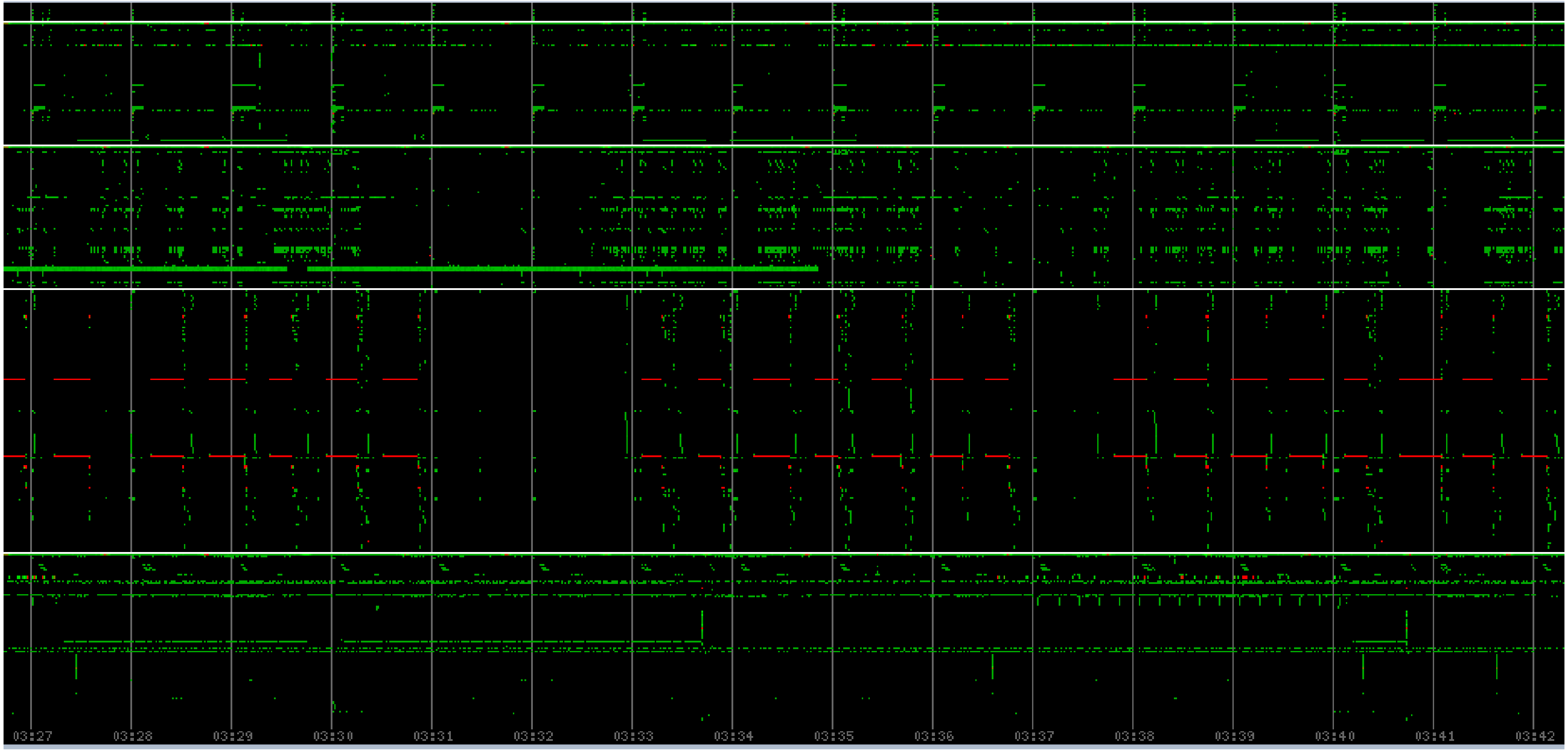
Technical Hurdles

- Log Message Fragmentation
  - When a log message gets broken up randomly into fragments, HELO recognizes it as a new log message and creates a new Template for it.
  - Causes highly increased number of Templates in the library and heavily bloats classification time.
  - Randomized fragments represent as unique events and give false positives of Type A events.
  - Solution: HELO automatic Template deactivation.
  - Correct solution: Fix source of log fragmentation.

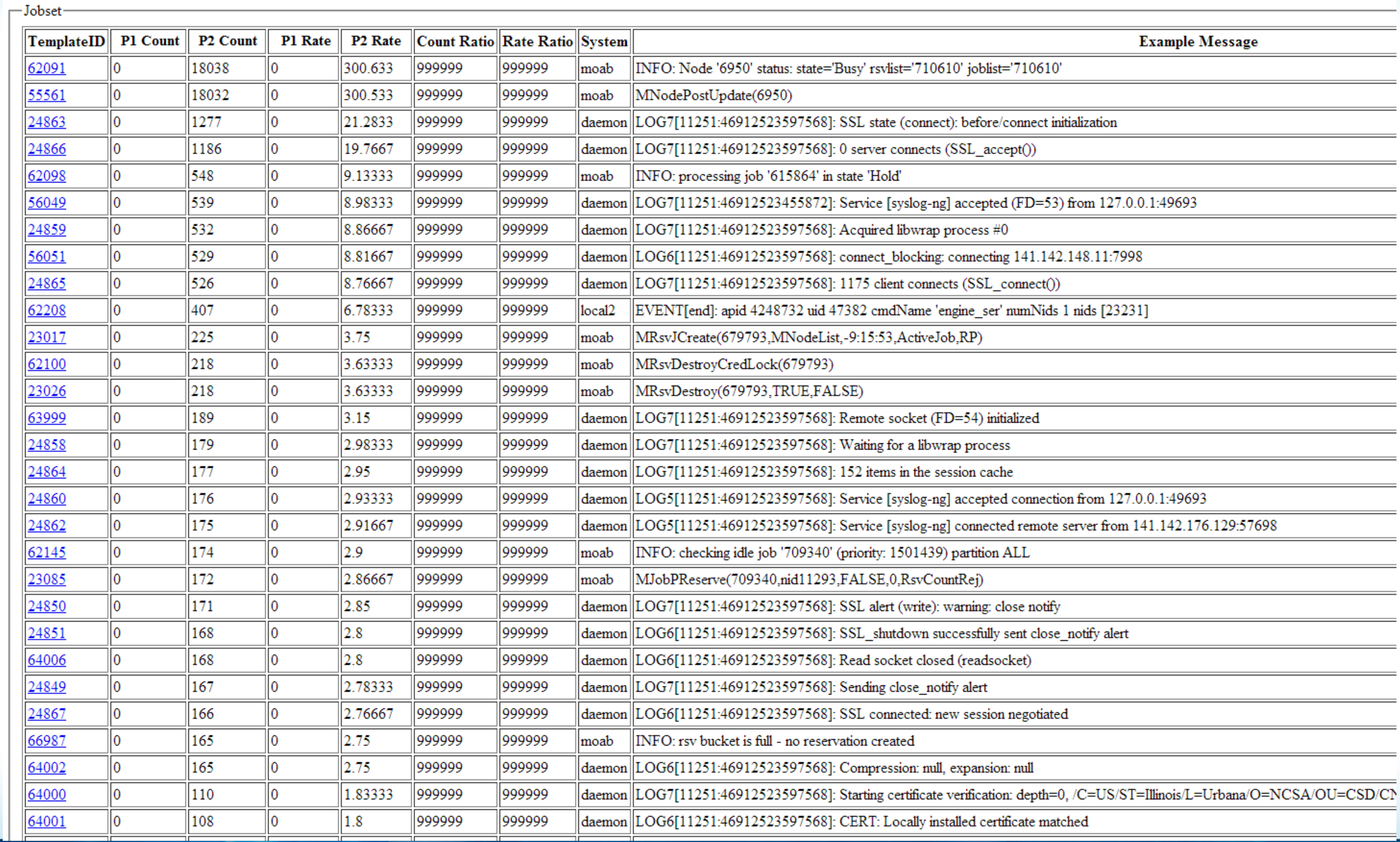
Technical Hurdles

- Fluctuating ‘Normal’
  - System upgrades, new or upgraded software, changes in logging levels, configuration changes, and many other things can all impact what shows up in the log streams.
  - Comparing logs from pre and post change will show differences, but results can be false leads.
  - Other job mix and shared resource contenetion also play a role.
  - Solution: Choose jobs or time periods that are temporally proximal. Lower probability that unrelated things will have greater variations.

Example


03:27
03:28
03:29
03:30
03:31
03:32
03:33
03:34
03:35
03:36
03:37
03:38
03:39
03:40
03:41
03:42

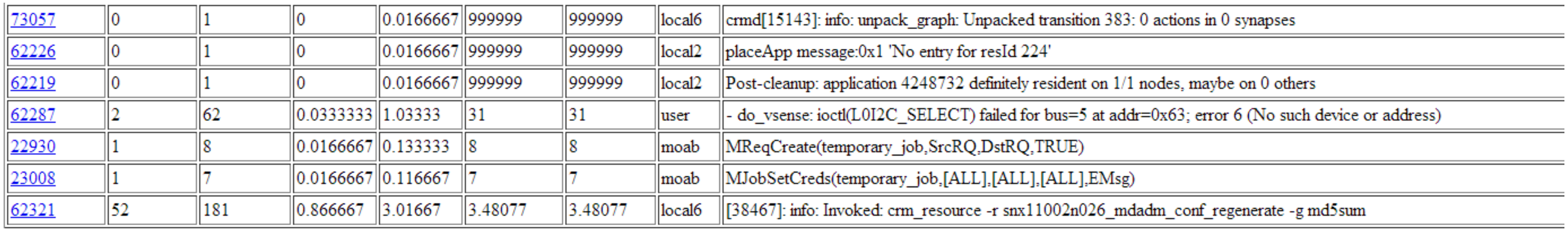
Example



Jobset

TemplateID	P1 Count	P2 Count	P1 Rate	P2 Rate	Count Ratio	Rate Ratio	System	Example Message
62091	0	18038	0	300.633	999999	999999	moab	INFO: Node '6950' status: state='Busy' rsvlist='710610' joblist='710610'
55561	0	18032	0	300.533	999999	999999	moab	MNodePostUpdate(6950)
24863	0	1277	0	21.2833	999999	999999	daemon	LOG7[11251:46912523597568]: SSL state (connect): before/connect initialization
24866	0	1186	0	19.7667	999999	999999	daemon	LOG7[11251:46912523597568]: 0 server connects (SSL_accept())
62098	0	548	0	9.13333	999999	999999	moab	INFO: processing job '615864' in state 'Hold'
56049	0	539	0	8.98333	999999	999999	daemon	LOG7[11251:46912523455872]: Service [syslog-ng] accepted (FD=53) from 127.0.0.1:49693
24859	0	532	0	8.86667	999999	999999	daemon	LOG7[11251:46912523597568]: Acquired libwrap process #0
56051	0	529	0	8.81667	999999	999999	daemon	LOG6[11251:46912523597568]: connect_blocking: connecting 141.142.148.11:7998
24865	0	526	0	8.76667	999999	999999	daemon	LOG7[11251:46912523597568]: 1175 client connects (SSL_connect())
62208	0	407	0	6.78333	999999	999999	local2	EVENT[end]: apid 4248732 uid 47382 cmdName 'engine_ser' numNids 1 nids [23231]
23017	0	225	0	3.75	999999	999999	moab	MRsvJCreate(679793,MNodeList,-9:15:53,ActiveJob,RP)
62100	0	218	0	3.63333	999999	999999	moab	MRsvDestroyCredLock(679793)
23026	0	218	0	3.63333	999999	999999	moab	MRsvDestroy(679793,TRUE,FALSE)
63999	0	189	0	3.15	999999	999999	daemon	LOG7[11251:46912523597568]: Remote socket (FD=54) initialized
24858	0	179	0	2.98333	999999	999999	daemon	LOG7[11251:46912523597568]: Waiting for a libwrap process
24864	0	177	0	2.95	999999	999999	daemon	LOG7[11251:46912523597568]: 152 items in the session cache
24860	0	176	0	2.93333	999999	999999	daemon	LOG5[11251:46912523597568]: Service [syslog-ng] accepted connection from 127.0.0.1:49693
24862	0	175	0	2.91667	999999	999999	daemon	LOG5[11251:46912523597568]: Service [syslog-ng] connected remote server from 141.142.176.129:57698
62145	0	174	0	2.9	999999	999999	moab	INFO: checking idle job '709340' (priority: 1501439) partition ALL
23085	0	172	0	2.86667	999999	999999	moab	MJobPReserve(709340,nid11293,FALSE,0,RsvCountRej)
24850	0	171	0	2.85	999999	999999	daemon	LOG7[11251:46912523597568]: SSL alert (write): warning: close notify
24851	0	168	0	2.8	999999	999999	daemon	LOG6[11251:46912523597568]: SSL_shutdown successfully sent close_notify alert
64006	0	168	0	2.8	999999	999999	daemon	LOG6[11251:46912523597568]: Read socket closed (readsocket)
24849	0	167	0	2.78333	999999	999999	daemon	LOG7[11251:46912523597568]: Sending close_notify alert
24867	0	166	0	2.76667	999999	999999	daemon	LOG6[11251:46912523597568]: SSL connected: new session negotiated
66987	0	165	0	2.75	999999	999999	moab	INFO: rsv bucket is full - no reservation created
64002	0	165	0	2.75	999999	999999	daemon	LOG6[11251:46912523597568]: Compression: null, expansion: null
64000	0	110	0	1.83333	999999	999999	daemon	LOG7[11251:46912523597568]: Starting certificate verification: depth=0, /C=US/ST=Illinois/L=Urbana/O=NCSA/OU=CSD/CN
64001	0	108	0	1.8	999999	999999	daemon	LOG6[11251:46912523597568]: CERT: Locally installed certificate matched

Example



73057	0	1	0	0.0166667	999999	999999	local6	crmd[15143]: info: unpack_graph: Unpacked transition 383: 0 actions in 0 synapses
62226	0	1	0	0.0166667	999999	999999	local2	placeApp message:0x1 'No entry for resId 224'
62219	0	1	0	0.0166667	999999	999999	local2	Post-cleanup: application 4248732 definitely resident on 1/1 nodes, maybe on 0 others
62287	2	62	0.0333333	1.03333	31	31	user	- do_vsense: ioctl(L0I2C_SELECT) failed for bus=5 at addr=0x63; error 6 (No such device or address)
22930	1	8	0.0166667	0.133333	8	8	moab	MReqCreate(temporary_job,SrcRQ,DstRQ,TRUE)
23008	1	7	0.0166667	0.116667	7	7	moab	MJobSetCreds(temporary_job,[ALL],[ALL],[ALL],EMsg)
62321	52	181	0.866667	3.01667	3.48077	3.48077	local6	[38467]: info: Invoked: crm_resource -r snx11002n026_mdadm_conf_regenerate -g md5sum

Fin

- Joshi Fullop (fullop@illinois.edu)
- Rob Sisneros (sisneros@illinois.edu)