Log Analysis and Iltop
LUG 2011

John Hammond
TACC
The University of Texas at Austin

jhammond@tacc.utexas.edu
Collaborators

- Tommy Minyard, TACC
- James C. Browne, U.T. C.S.
- Edward Chuah, Howie Kuo, Gary Khoon, William Tjhi, IHPC Singapore
Part 1: Log Analysis

Log analysis is for the birds.

- Kent B.

System logs are a kind of cognitive black hole. It's easy to pour vast amounts of time and energy in to staring at them, but then have nothing ever seem to come out.

-A.
Why bother with logs?

- Good trace (\texttt{printk()}) better than debugger, SystemTap, bad trace, ...
- They're there: easy to collect with no added infrastructure.
Rational message encoding

In the kernel, modify printk() by inserting call to a new function, rat_printk().

Every call to printk(),

```c
printk(KERN_ERR "device failed: Error %d: %s
", errno, errmsg);
```

would produce the normal message:

```
Mar 10 15:09:30 host1 kernel: device failed: Error 157: Not a toaster
```

along with an unformatted version of the same (line breaks inserted for clarity):

```
time:1268255370\0
host:host1\0
programe:kernel\0
0:<3>Print failed: Error %d: %s\n\0
1:157\0
2:Not a toaster\0\0
```
Transport parallel to `printk()`

- Store in ring buffer in kernel.
- Read from `/proc/rat-kmsg` (message oriented)
- Run rat-syslogd to add timestamp, hostname, store, forward, ...
Comparison 1

Jul 8 12:27:52 i115-312 kernel: test_sine.x[27673]: segfault \
at 00002aaa9e094c88 rip 00002b92854db19e rsp 00007fff25bc0b40 \error 4
LustreError: 11-0: an error occurred while communicating with 129.114.97.41@o2ib. The ost_write operation failed with -107
Comparison 3

LustreError: 1943:0:(osc_request.c:1143:can_merge_pages()) is it ok to have flags 0xc20 and 0x420 in the same brw?

0:%sLustreError: %d:%d:(%s:%d:%s()) is it ok to have flags 0x%x\ and 0x%x in the same brw?
1:<3>
2:1943
3:0
4:osc_request.c
5:1143
6:can_merge_pages
7:3104
8:1056
Apr 12 01:07:04 i115-409 kernel: Lustre: 3570: (client.c:1476: ptlrpc_expire_one_request()) @@ Request x136339563010 0532 sent from scratch-OST009a-osc-ffffff810877104c00 to NID 129.114.97.67@...

$ rat-log-find -msg 'Apr 12 01:07:04 i115-409 kernel: Lustre: 3570'
time:1302588424
host:i115-409
jobid:1902176
prog:kernel
0:%sLustre: %d:(%s:%d:%s()) @@ Request x%Lu sent from %s to NID ...
1:<4>
2:3570
3:client.c
4:1476
5:ptlrpc_expire_one_request
6:1363395630100532
7:scratch-OST009a-osc-ffffff810877104c00
8:129.114.97.67@o2ib
...
40:0
Apr 12 01:07:04 i115-409 kernel: Lustre: 3570:(client.c:1476:\
ptlrpc_expire_one_request()) @@ Request x136339563010 0532 \sent from scratch-OST009a-osc-ffff810877104c00 to NID 129.114.97.67\@02ib 7s ago has timed out (7s prior to deadline).
Apr 12 01:07:04 i115-409 kernel:  req@ffff8103ee3a7000 \
x1363395630100532/t0 o101->scratch-OST009a_UUID@\129.114.97.67@02ib:28/4 lens 576/544 e 0 to 1 dl 1302588424 \
ref 1 fl Rpc:/0/0 rc 0/0

$ rat-log-find -msg 'Apr 12 01:07:04 i115-409' -print 0
%sLustre: %d:(%s:%d:%s()) @@ Request x%Lu sent from %s to NID %s\%lus ago has %s (%lds prior to deadline).\n req%p x%Lu/t%Ld \o%d->%s@%s:%d/%d lens %d/%d e %d to %d dl %ld ref %d fl %s:%s%s%s\%s%s%s%s/%x/%x rc %d/%d

$ rat-log-find -msg 'Apr 12 01:07:04 i115-409' -print 8
129.114.97.67@02ib

$ rat-log-find -msg 'Apr 12 01:07:04 i115-409' -print 15
101
Part 2: TACC lltop

- Lustre load monitor with batch (SGE,...) scheduler integration
- Point-and-shoot operation.
- Helps decide where to send the goons.
<table>
<thead>
<tr>
<th>JOBID</th>
<th>WR_MB</th>
<th>RD_MB</th>
<th>REQS</th>
<th>OWNER</th>
<th>WORKDIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1901914</td>
<td>248</td>
<td>0</td>
<td>641</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>1901919</td>
<td>180</td>
<td>0</td>
<td>727</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>1901959</td>
<td>143</td>
<td>0</td>
<td>622</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>ilogin3</td>
<td>90</td>
<td>33</td>
<td>2463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1901923</td>
<td>87</td>
<td>0</td>
<td>524</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>1902767</td>
<td>86</td>
<td>0</td>
<td>276</td>
<td>alice</td>
<td>/work/01117/alice/CC_1p</td>
</tr>
<tr>
<td>1901957</td>
<td>80</td>
<td>0</td>
<td>1683</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>1902703</td>
<td>51</td>
<td>0</td>
<td>104</td>
<td>edward</td>
<td>/work/01682/edward/H20_flu</td>
</tr>
<tr>
<td>1901838</td>
<td>44</td>
<td>0</td>
<td>88</td>
<td>samuel</td>
<td>/work/00934/samuel/4_1</td>
</tr>
<tr>
<td>1902684</td>
<td>10</td>
<td>0</td>
<td>23</td>
<td>mary</td>
<td>/work/00491/mary/sims/dma</td>
</tr>
<tr>
<td>1901166</td>
<td>9</td>
<td>0</td>
<td>23</td>
<td>james</td>
<td>/work/01516/james/big</td>
</tr>
<tr>
<td>1900672</td>
<td>8</td>
<td>0</td>
<td>19</td>
<td>james</td>
<td>/work/01516/james/small</td>
</tr>
<tr>
<td>1902252</td>
<td>6</td>
<td>0</td>
<td>66</td>
<td>cora</td>
<td>/work/01333/cora/123/prod</td>
</tr>
<tr>
<td>1901738</td>
<td>6</td>
<td>0</td>
<td>41</td>
<td>annie</td>
<td>/work/00287/annie/fg20</td>
</tr>
<tr>
<td>1901917</td>
<td>5</td>
<td>0</td>
<td>2340</td>
<td>nellie</td>
<td>/work/01251/nellie/XYZ20/Lu</td>
</tr>
<tr>
<td>1901305</td>
<td>4</td>
<td>0</td>
<td>29</td>
<td>minnie</td>
<td>/work/00914/minnie/MUT/prod</td>
</tr>
<tr>
<td>1902646</td>
<td>4</td>
<td>19</td>
<td>52</td>
<td>frank</td>
<td>/work/01640/frank/CORAL/R1</td>
</tr>
<tr>
<td>1900742</td>
<td>3</td>
<td>0</td>
<td>35</td>
<td>bertha</td>
<td>/work/01246/bertha/amber/cl</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
lltop architecture
Design goals.

• Single executable to install on servers.

• Invocation and transport by ssh.

• Hookable
  – LNET-address to hostname conversion.
  – Hookable hostname to jobid conversion.
  – FS name to server list.
Click Here

to download the GUI of the future!

Here Today!

https://github.com/jhammond/lltop
Part 3: Lustre /proc

• Mostly very good.
• Needs some attention.
  – Some are just wrong.
  – Some are only partially wrong.
  – What is a reg?
• Force multiplier.
  – Enable toolmakers to add value without mucking about in Lustre source.
Wish list 1.

• Make it easier for zero-conf tools to associate servers, /proc directories with mount points.
  – What if they mount two scratches?
  – Appending “_UUID” does not make it so.

• Make “Bytes on the wire” stats available in one file.
Wish list 2.

- Keep /proc.

- Text only.

- No special tools please.

- Strict “No ASCII art policy.”