Lessons Learned From A Very Unusual Year

Kevin Buterbaugh
Kevin.Buterbaugh@vanderbilt.edu

IBM GPFS User Group Meeting
SC18 - Dallas, TX
November 11th, 2018

ACCRE
Advanced Computing Center for Research & Education

SC18
Dallas, TX | hpc inspires.

How To Survive 2 Filesystem Corruption Incidents, Multiple Hardware Failures, Usage - Both From a Performance and Capacity Perspective - Exploding, and Still Upgrade Your Cluster to GPFS 5 - All in the Same Year, While Retaining Your Sanity - If Not All Your Hair

Kevin Buterbaugh
Kevin.Buterbaugh@vanderbilt.edu

IBM GPFS User Group Meeting
SC18 - Dallas, TX
November 11th, 2018

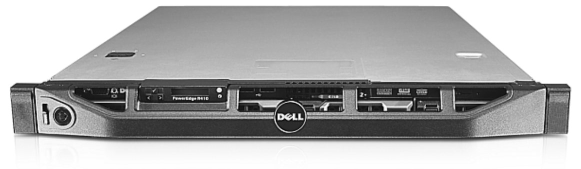
ACCRE
Advanced Computing Center
for Research & Education

SC18
Dallas, TX | hpc inspires.

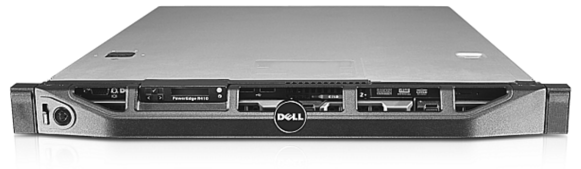
Some Background

- The Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University operates a ~10,000 core Linux cluster used by over 1,000 researchers at the University and Medical Center.
- We have been running GPFS in production since August of 2005.
- The first version of GPFS we installed was 2.3.
- We're currently migrating from 4.2.3-x to 5.0.2.x.
- We currently have two GPFS filesystems:
  - /gpfs22 (home directories): 25 TB (usable)
  - /gpfs23 (/scratch and /data): 1.2 PB (usable)

GPFS Configuration



2 GPFS Cluster / Filesystem Managers
Also serve as pmcollector and GPFS GUI nodes



8 GPFS NSD Servers - 2U Servers with 2 quad-core Intel CPUs, 64GB RAM, dual-port 8 Gb Fibre Channel, 10 Gbit Ethernet

4 QLogic SANbox 5800 8 Gb
Fibre Channel Switches

All NSD servers and Storage
Arrays are connected to both SAN
switch stacks for redundancy

1 Infortrend Eonstor Storage
Array for storing metadata

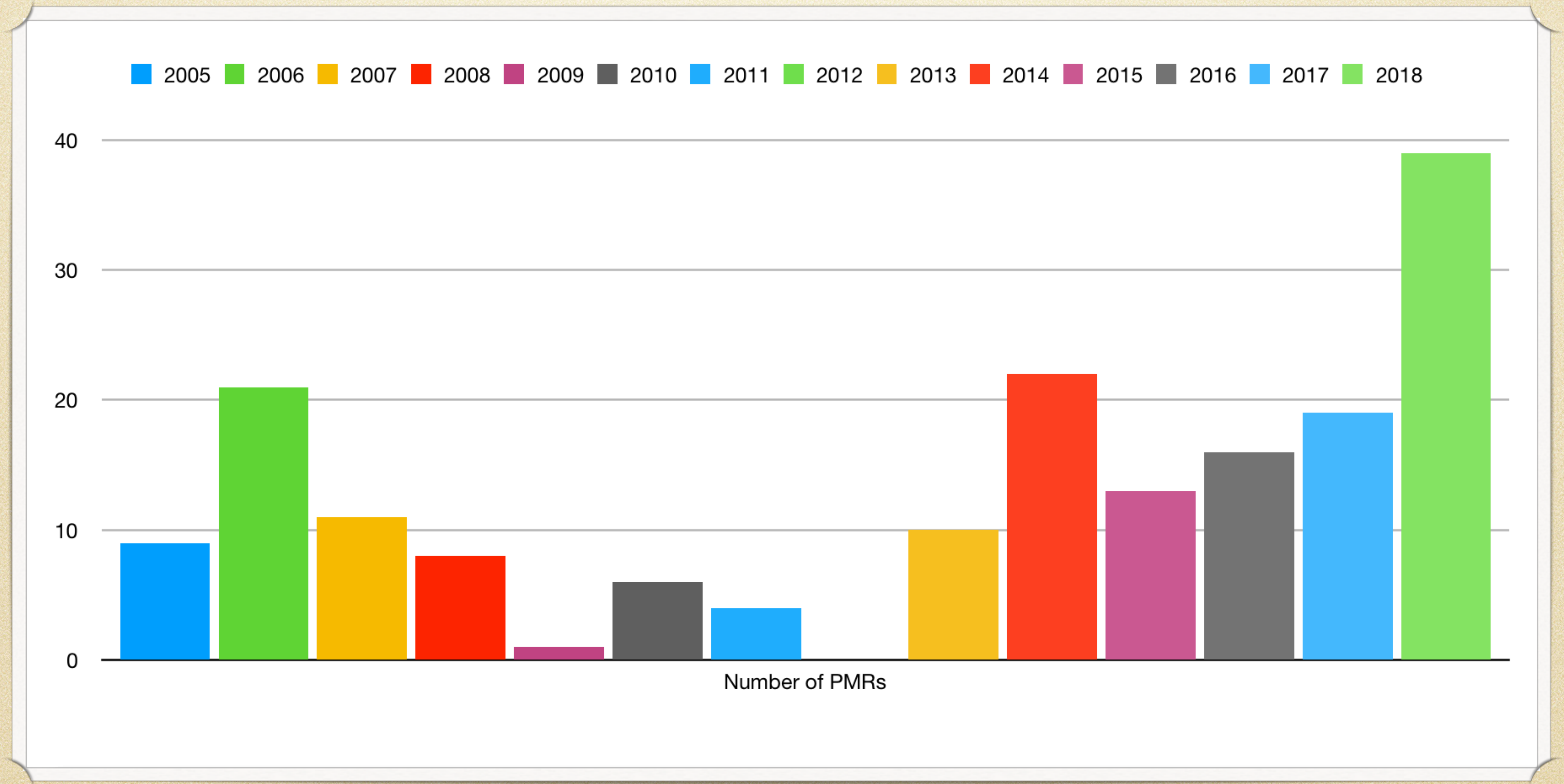
dual redundant 8 Gb FC controllers with 16 GB cache per controller
and 4 200 GB and 12 800 GB SSDs configured with RAID 1

10 Infortrend Eonstor Storage Arrays
and 6 JBODs for storing data
9 Eonstors Arrays have 24 drive bays,
1 has 60 drive bays
JBODs have 16 drive bays

dual redundant 8 Gb FC controllers
with 4 GB cache per controller and 3,
4, or 8 TB SATA or SAS hard drives
configured with RAID 6

Before we get into 2018 itself

- Number of PMR's / Tickets opened per year:



You can shout “fire” if there is one!

- On March 31st, 2016, a construction worker who had been given incorrect blueprints jackhammered into the main electrical feed to our data center, resulting in an electrical fire in the data center. Most importantly, he survived.

- The chemical suppressant did its job.

- Unfortunately, every piece of equipment in the room had that chemical - plus soot and smoke particles - on them and *in* them and had to be professionally cleaned.

- On the plus side, we were without air conditioning in the data center for over 1 hour, the temperature in the room exceeded 100 degrees … and GPFS never went down!

Sunday, January 24th, 9:04 AM

- I received a text message from the on call person asking me to look at a Help Desk ticket. It appeared the /gpfs23 filesystem had unmounted across the cluster.

- I quickly confirmed that the filesystem was unmounted and couldn't be remounted. We had long waiters and the filesystem manager for gpfs23 was unresponsive. In addition, mmlsdisk showed two disks down.

The Initial Suspect

- The gpfs23 filesystem has 3 pools: 1) system (metadata only), 2) gpfs23data, and 3) gpfs23capacity.
- The previous week I had added two new disks to the system pool and kicked off an mmrestripefs of that pool.
- On Friday evening a cron job which migrates data between the gpfs23data and gpfs23capacity pools also kicked off (mmapplypolicy).
- A conflict between the mmrestripefs and mmapplypolicy that somehow corrupted data was suspected.

And so began a very long week

- IBM instructed me to run an mmfsck, but it hung up.

- The two down disks were the two metadata disks I added the prior week, so a problem with the mmadddisk was suspected.

- However, the mmadddisk had completed successfully and I had subsequently kicked off the mmrestripefs (I *always* do those two steps separately).

- I won't include all the details (a PDF of the web page for this PMR is 98 pages long!), but it literally took 8 days of around the clock work to bring the two disks up, get an mmfsck to complete, and identify ~300 corrupted files (out of 300 million).

The root cause of the problem was?

- Nobody knows for sure.
- IBM suspected a hardware problem during the mmadddisk / mmrestripefs caused both copies of the metadata for those ~300 files to be corrupted.
- However, I uncovered e-mails from our tape backup vendor pointing out that their software was reporting read errors on those same files weeks before I had added the disks!
- And there was a point in the fall of 2017 when we were running a GPFS version known to have a silent data corruption bug (subsequently fixed).

Lessons Learned

- Don’t ignore *any* signs of filesystem corruption (duh!). The initial signs of trouble happened over the holidays and got lost in the noise / lack of paying attention.

- Having multiple filesystems is a good idea! While users were not happy that the majority of their data was inaccessible for a little over a week, the fact that they could still log on to gateways and work with the files in their home directories really helped keep the frustration level down.

- The fact that IBM could provide 24x7 support for this incident by rolling us between the teams in the United States, Europe, and China was great! The fact that we could only roll the SysAdmin working with them from Kevin to Kevin to Kevin … not so great!

Monday, April 2nd, 1:43 PM

- While investigating a non-GPFS related Help Desk ticket, I noticed the following in /var/log/messages:
- Error=MMFS_FSSTRUCIT, ID =0x94B1F045, Tag=-6070658: Invalid disk data structure. Error code 1124. Volume gpfs22. Sense Data
- Looking at other logs revealed that these errors have been occurring for several weeks! Sigh…
- However, if GPFS still did not log to the system log we wouldn't have found this issue as soon as we did.

This one was straightforward

- …if not painless.
- Error 124 is a bad directory entry and requires an offline mmfsck to fix.
- Since this was on our gpfs22 (/home) filesystem, we scheduled a downtime over the weekend and ran the mmfsk, which fixed the problem.

Lessons Learned

- Don’t forget that you have to actively look for filesystem corruption!

- GPFS callbacks are your friend! After this incident, ACCRE management prioritized my writing callbacks for this and other issues (ex: GPFS disk down) and integrating them with our Help Desk ticketing system as appropriate.

- Having multiple filesystems is a good idea! While users were not happy that their home directories were unavailable while we ran the mmfsck, the fact we ran the mmfsck over the weekend and that most running jobs were not killed (since they were primarily accessing data on the gpfs23 filesystem) helped keep the frustration level down.

Wednesday, May 30th, 8:50 PM

- We received two e-mail alerts:
  - One of our storage arrays reported a controller down.
  - One of the GPFS callbacks we had written reported a disk down in our gpfs23 filesystem … and on the same storage array as the down controller.
- The storage arrays all have dual-redundant active-active controllers. However, of the 3 LUNs assigned to the now failed controller, only 2 of them failed over to the other controller.
- More troubling, the LUN showed up in the management interface for the storage array with a size of 0 bytes.

This was not a GPFS issue

- There were two issues with the storage array:
  - A failed controller.
  - A LUN that should've been ~22 TB in size was now 0 bytes.
- Infortrend support was able to recover the failed controller, but they never were able to fix the problem with the 0 byte LUN.
- We used the mmfileid command to identify the files with one or more blocks on the affected LUN and restored those from tape.

Lessons Learned

- Past performance is no predictor of future performance. Infortrend arrays had been rock-solid for us for ~15 years.
- While no root cause of either issue with the array was every definitively identified, we suspect a combination of:
  - Firmware bugs.
  - Flakiness caused by the fire and subsequent cleaning.
- GPFS callbacks are your friend!
- Not only is tape not dead, tape backups can save your bacon.

Monday, September 1st, 4:30 AM

- I arrived at work to upgrade the RAM / Cache in our older storage arrays.
- The backstory:
  - Earlier in the summer we began receiving complaints about sluggishness in our GPFS filesystems.
  - After spending a significant amount of time tracking this down, including working with IBM support, we determined (from "mmdiag —iohist") that at certain times the actual I/O's to disk were taking 1 - 10 (or more) *seconds*!
  - From that information we quickly determined that on our older storage arrays the 2 GB cache per controller was being overrun.

Monday, September 1st, 4:30 AM

- The backstory (continued):
  - The older arrays support a max of 4 GB RAM per controller.
  - The storage arrays were EOL'd by the vendor, so we bought 3rd party RAM and tested upgrading the RAM live - i.e. one controller at a time - in an array that was part of our test cluster. This worked.
  - We then took that same RAM and upgraded one of our production cluster storage arrays live. This also worked.
  - We then ordered enough of the same 3rd party RAM to upgrade the rest of the older storage arrays … which brings us back to September 1st at o'dark thirty AM…

Monday, September 1st, 4:30 AM

- We replaced the RAM on one controller in another of the production cluster storage arrays … and the controller failed to boot!
- Thinking this might possibly be just a flaky controller, we did the same thing with one controller in a second storage array … and it also failed to boot!
- We opened a ticket with Infortrend support, who informed us that having different sized DIMMs in the controllers was not supported.
- We asked them why it worked - twice - and their response was that having different sized DIMMs in the controllers was not supported.
- However, they did help us recover the down controllers.

Lessons Learned

- Make sure that the maintenance you want to perform is supported by the vendor.
- Obviously, just because it works in your tests doesn’t mean it’s going to work when you really need it to.
- Never forget Murphy’s Law!

October and November - Periodically

- Users would report:
  - Simple commands (ex: cd, ls, vim) hanging.
  - Logging in to the cluster hanging.
- We could duplicate the problem.
- A look in the GPFS logs showed that each incident was preceded by node expels and a message similar to:
- 2018-11-06_14:58:48.554-0600: [I] Recovery: gpfs22, delay 79 sec. for safe recovery.

But why were nodes being expelled?

- The first thing we checked was the network, both:
  - From the node to the top of rack switch, and
  - From the top of rack switch to our central switch stack.
  - But both were OK.
- Next we checked for resource exhaustion (CPU / RAM) on the node itself … also OK.
- Bad RAM? Bad CPU(s)? No.
- The problem turned out to be voltage fluctuations causing nodes to crash!

Lessons Learned

- Sometimes "GPFS problems" aren't GPFS problems.
- GPFS callbacks are your friend! The callback I had written for a "node leave" event was our first lead in tracking down this problem.
- While the network was a logical first thing to look at - never assume.
- Never rule anything out until you're 100% sure you can rule it out.
- Remember Murphy's Law!

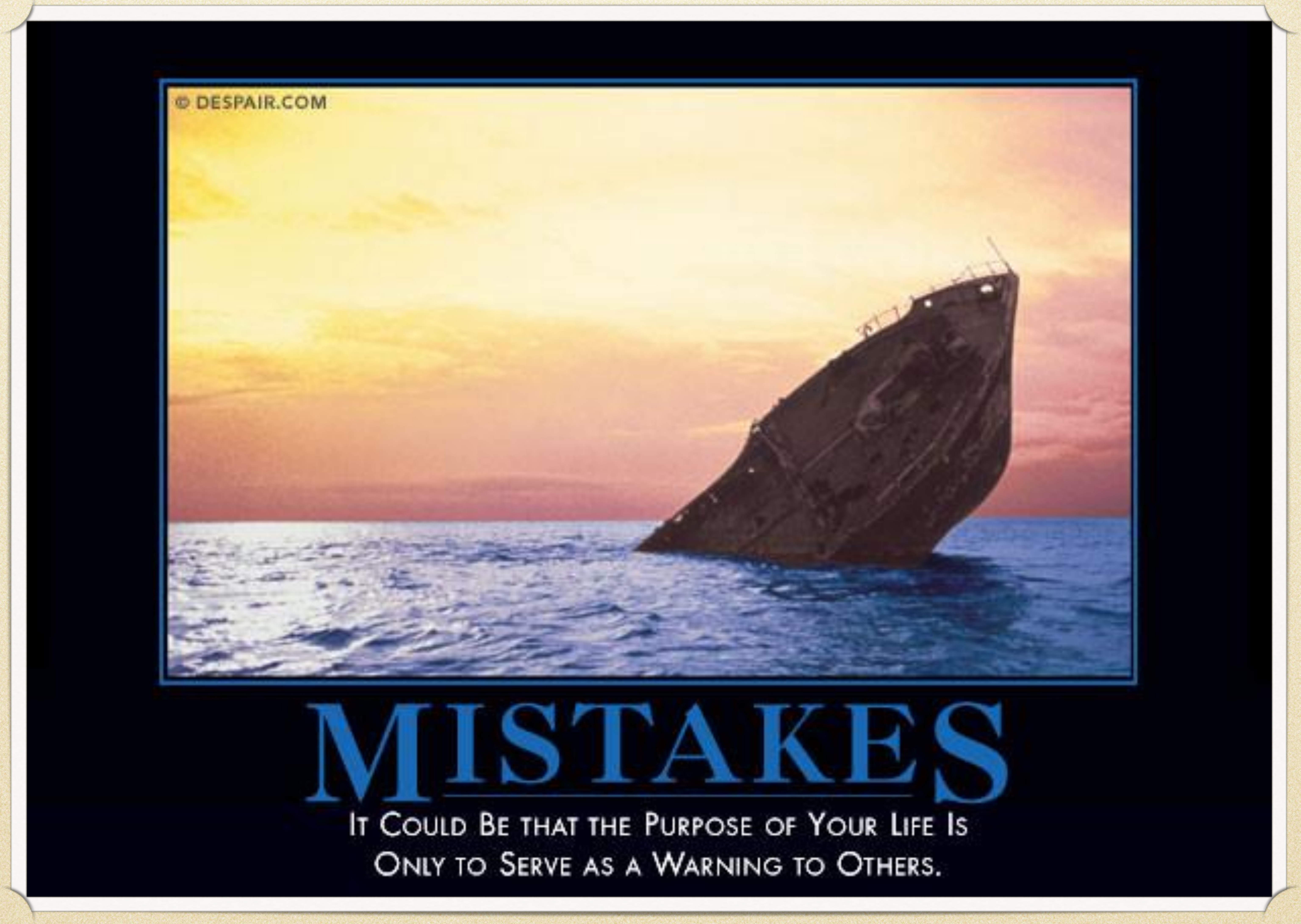
Miscellaneous Notes

- To help with our performance and capacity issues we have purchased:
  - A 24 bay Eonstor with 128 GB RAM per controller and 10 SSD’s to be used for metadata on our new GPFS 5 filesystem.
  - A 60 bay JBOD with 60 12 TB drives (note that a 12 TB drive isn’t even 11 TB!) SAS connected to the Eonstor.
  - We are in the process of testing various configs for metadata and will go to production with this hardware in either December or January.
- We plan to purchase a second such Eonstor / JBOD in 2019 and then life-cycle all of our old storage arrays.

Overall Lessons Learned

- IBM Support really is great. If IBM tells you something, believe them unless you’ve got good evidence to the contrary.
- The GPFS mailing list is great, as are these User Group meetings.
- The IBM DeveloperWorks wiki pages related to GPFS contain a wealth of useful information.
- GPFS callbacks are your friend. Execute “man mmlscallback” and write callbacks for anything that might apply to you (and I’ll be happy to share any of the ones I’ve written with anyone interested).
- If at all possible, have more than one staff member experienced with GPFS (this is something we’re still working on at Vanderbilt).

We all have a role to play!


© DESPAIR.COM
MISTAKES
It Could Be that the Purpose of Your Life Is
Only to Serve as a Warning to Others.

Questions / Comments?