Nils Haustein | Executive IT Specialist | EMEA Storage Competence Center

IBM Spectrum Scale | Automation of Storage Services

Agenda

- **Overview and challenges**
- Implementation guidance
- Hints and tips

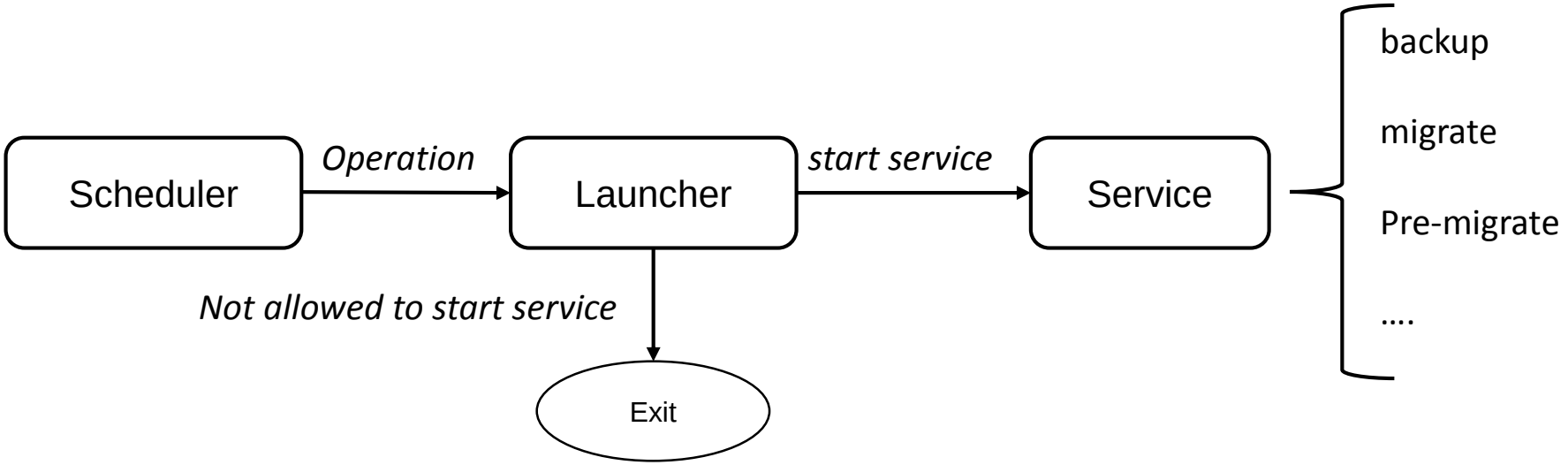
Overview

- Spectrum Scale storage service are background tasks such as:
  - Backup
  - Migration & pre-migration
  - Scale out Backup and Recovery
  - Snapshots
  - List policies generating certain statistics
- Storage service are typically scheduled and run unattended
  - There are various challenges with running storage services in a cluster
- This presentation discusses the challenges and gives guidance for solutions

Challenges with scheduling storage services

- Node selection for starting the storage service
  - Storage services are started on one node and run on all or a subset of nodes
  - On which cluster node should the storage service be launched?
  - What happens if the node selected to start the service is down?
- Determining the status
  - Storage service may require certain resources to be operational
  - Such as the nodes to execute the service, file systems, etc.
- Monitoring the result of a storage service
  - How to determine if a storage service operation was successful or not
  - Notification about failures of storage service operation is desireable
- Investigating failures of a storage service operation
  - Consistent logging of storage services operations

High level concept



Scheduler	Launcher	Storage service
Starts the operation on all nodes that are allowed to run it Can be implemented with cron or systemd timers	Script implemented on all nodes that are allowed to run it Determines if the nodes is allowed to run it, if not exit Checks node and cluster status and launches the service	Script, implemented on all nodes that are allowed to run it Run the storage service operation

Agenda

Overview and challenges

Implementation guidance

Hints and tips

Launcher 1/3

- Launcher script is implemented on all nodes with manager role
- Determines if this node is cluster manager
  - Cluster manager is a unique role within the cluster
  - Cluster manager is available as long as the cluster is online
  - Nodes that are not cluster manager exit after this check

```
# determine local node name
localNode=$(mmlsnode -N localhost | cut -d'.' -f1)
# determine the cluster manager name
clusterMgr=$(mmlsmgr -c | sed 's|.*(||' | sed 's|)||')

# check if the local node is the CM, if not exit 0
if [ "$localNode" != "$clusterMgr" ];
then
  echo "INFO: this node ($localNode) is not cluster manager, exit."
  exit 0
fi
```

Launcher 2/3

- Determines the status of the cluster
  - Status of the cluster
  - Status of nodes required to run the service
  - Status of the file system subject for the storage service

```
# check if file system is mounted on the local node
mounted=0
mounted=$(mmlsmount $fsName -L | grep "$localNode" | wc -l)
if (( mounted == 0 ));
then
  echo "ERROR: file system $fsName is not mounted on node, exit."
  exit 2
fi
```

Launcher 3/3

- Assigns the log-file for the operation

```
# operation is the first argument passed to the launcher
op=$1
# directory for log files
logDir="/var/adm/ras/storageservice"
# current date will be part of the log file name
curDate="$(date +%Y%m%d%H%M%S)“

# assing logfile name
logF=$logDir"/"$op"_"$curDate".log"
```

- Starts the storage service script with the required scope (e.g. file system)
  - Optionally it may defer the operation to another available node

```
eval $cmd >> $logF 2>&1
rc=$? # analyze return code and notify the admin when required…
```

Storage service

- Storage service script performs the operation with the required scope
  - For example: for snapshot-backup it creates a snapshot prior to the backup
    - And delete it afterwards
- There might be one script for each type of operation
  - Backup, migrate, sobar, etc.
- Console output of the service script is redirected to the log-file
- Each storage service script should implemented consistent return codes
- Storage service script is installed on all nodes with manager role
  - Optionally on other nodes where the operation is deferred to by the launcher

Scheduler

- Scheduler starts the launcher script on nodes with manager role
  - Passes the operation and the scope for the operation
  - Example with crontab launching backup

```
PATH=/usr/bin:/usr/sbin:/usr/lpp/mmfs/bin

00 06 00 00 00  /path-to-scripts/launcher.sh backup
```

- All nodes with manager role must have the same schedules active
- Can be implemented with cron, systemd timer-units or external scheduling

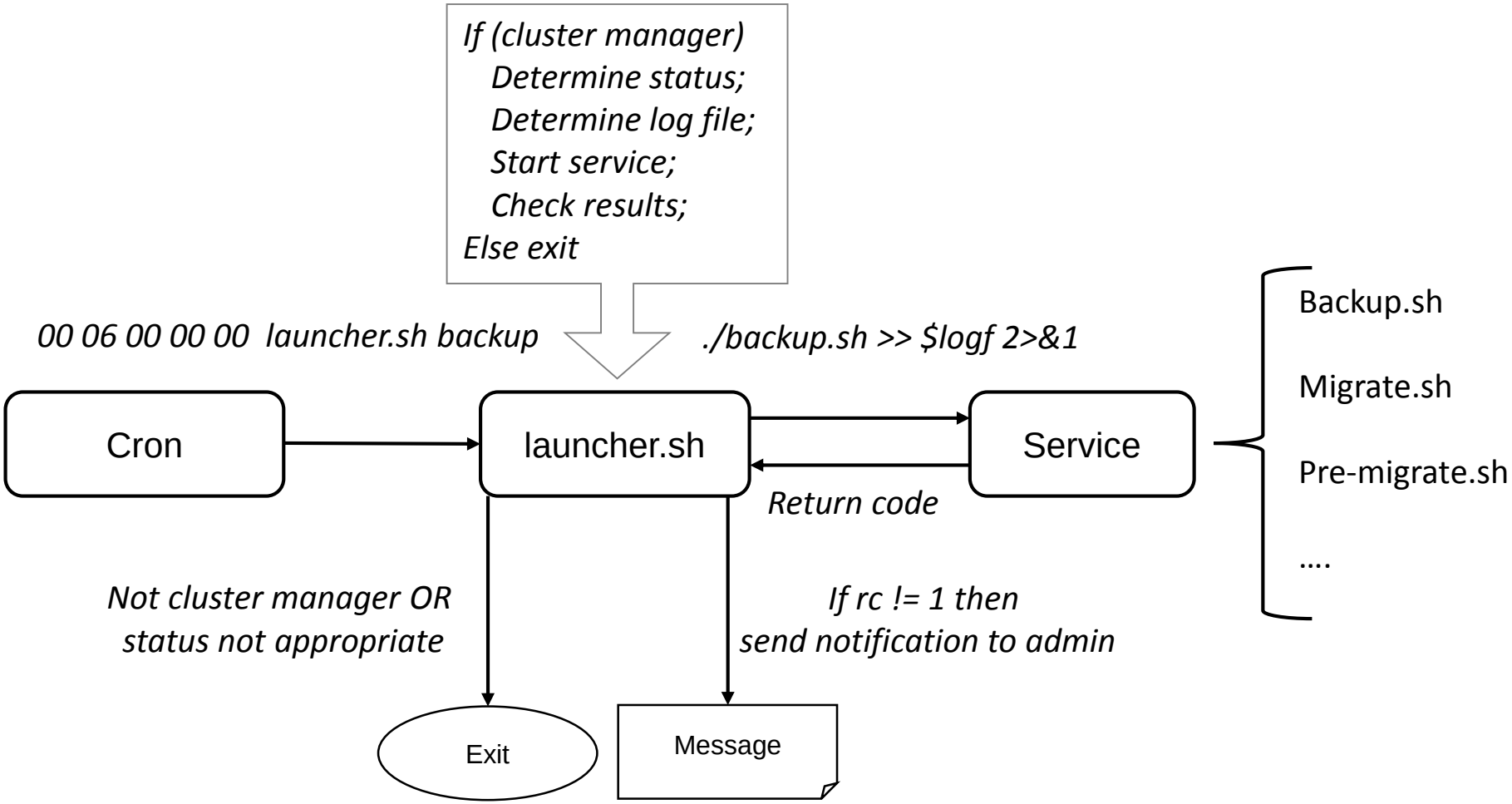
Logging

- Launcher determines the log-file name
  - Based on operation and scope
  - Example: `/var/adm/ras/backup_myfs_170220120000.log"`
- Launcher redirects output of service script to log-file
  - Example: `eval $cmd >> $logF 2>&1`
- Launcher may manage log-files
  - Number of version to be kept
  - Number of versions to be compressed
  - Can also leverage log-rotate function of the operating system

Monitoring

- Report (failure) results of storage services to storage admin
  - For example by using emails
- Can be done by launcher, because it knows:
  - Operation
  - Scope of operation
  - Return code of service script
  - Name of the log-file
- Not (yet) supported method: send custom event to the GUI
  - When GUI is configured for event notification it sends email or SNMP trap

Low level concept



Agenda

Overview and challenges

Implementation guidance

Hints and tips

Display log-files from all cluster nodes

- Assumes that log-files have consistent naming based on operation
  - such as /var/adm/ras/*operation_date.log*
- Assumes that certain tokens are used within the log-files
  - such as BACKUP or MIGRATE

- Processing:
  - Determine the most recent log-file for the operation based on consistent file names
  - Filter the most recent log-file based on pre-defined tokesn
  - Perform this operation on all relevant nodes using ssh (or mmdsh)

```
Last log for node g1_node1:
CHECK: started with operation backup on $(g1_node1)
CHECK: checking if this node is cluster manager
------------------------------------------------------------------------
Last log for node g1_node2:
CHECK: started with operation backup on $(g1_node2)
CHECK: checking if this node is cluster manager
CHECK: checking if this node has file system gpfs1 mounted
NORMAL-BACKUP: Starting mmbackup
NORMAL-BACKUP: Finished mmbackup (rc=0)
CHECK: command backup.sh finished with rc=0
```

Backup using mmbackup

- mmbackup creates different temporary files and stores it in directory (/tmp)
  - Policy result files
  - Shadow DB of mmbackup
  - Result files of comparison of Spectrum Protect inventory and shadow DB
- If the underlying file system is out of space mmbackup fails
- Directory for mmbackup files can be controlled with parameters –s and –g
  - if –s is set it also sets –g
  - Recommendation: set –s and –g to directory that has sufficient space
  - Practical approach: store mmbackup files in the Spectrum Scale file system
    - e.g. under /ibm/gpfs1./mmbackupworkdir
  - exclude this directory from backup using EXCLUDE statement in dsm.sys
- How much temp space do I need ?
  - mmbackup shadow DB format: link

Aligning mmbackup, client and server options

Align the mmbackup, Spectrum Protect client and server options

Define number of bytes transferred in a single backup transaction

`--max-backup-size` (N *) `TXNBYTELIMIT`

Define number of files transferred in a single backup transaction

`--max-backup-count`
`--max-expire-count`
(N *) `TXNGROUPMAX`

Define number of parallel backup processes

`--backup-threads`
`--expire-threads`
(* #Nodes *) `RESOURCE-UTILIZATION` `MAXSESSIONS` `MAXNUMMP`

Spectrum Scale **mmbackup** command line options

Spectrum Protect **B/A client** processing options

Spectrum Protect **Server** processing options

Migration and pre-migration

- Migration is invoked using mmapplypolicy
  - Consider using parameters –s and –g to control where these files are stored
- Pre-migration can be invoked with special threshold based policy

```
RULE 'premig' MIGRATE FROM POOL 'system' THRESHOLD (0,100,0) TO POOL 'ltfs'
```

- Alternatively create external pool script for pre-migration
  - copy sample script to `/usr/lpp/mmfs/samples/ilm/`**`mmpolicyRules-hsm.premig`**
  - Adjust the following line at the beginning:

```
$MigrateFormat = "%s %s -premigrate -filelist=%s";
$PremigrateFormat = "%s %s -premigrate -filelist=%s";
```

  - Define policy to perform pre-migration

```
define( is_managed,(MISC_ATTRIBUTES LIKE '%M%') )
RULE EXTERNAL POOL 'hsm' EXEC '/…/mmpolicyRules-hsm.premig' OPTS '-v'
RULE 'PreMig' MIGRATE FROM POOL 'system' TO POOL 'HSM' WHERE NOT
is_managed
```

Recommended reading

https://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102676

What to learn more about Spectrum Scale?

- Spectrum Scale Standard hands-on workshop
  - Learn about architecture & concepts, installation and configuration, ILM, CES, AFM, Backup, ….
  - https://academy.avnet.com/de/training/course/141611
- Spectrum Scale Advanced hands-on workshop
  - Learn about architecture & concepts, monitoring, configuration parameters, tools and problem determination ….
  - https://academy.avnet.com/de/training/course/136736

→We also offer customized workshops according to your needs !

References and Links

- Automation whitepaper
  https://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102676
- Gitlab project (IBM Internal)
  https://github.rtp.raleigh.ibm.com/nils_haustein-de/Spectrum-Scale-Automation
- mmbackup shadow DB format:
  https://www.ibm.com/support/knowledgecenter/STXKQY_4.2.2/com.ibm.spectrum.scale.v4r22.doc/bl1adv_recordformat.htm
- Peta-scale data protection with Spectrum Protect
  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Tivoli%20Storage%20Manager/page/Petascale%20Data%20Protection

Disclaimer

- This information is for **IBM Spectrum Scale user day 2017 use only**, publication beyond this scope is forbidden
- This information is provided on an "AS IS" basis without warranty of any kind, express or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. Some jurisdictions do not allow disclaimers of express or implied warranties in certain transactions; therefore, this statement may not apply to you.
- This information is provided for information purposes only as a high level overview of possible future products. PRODUCT SPECIFICATIONS, ANNOUNCE DATES, AND OTHER INOFORMATION CONTAINED HEREIN ARE SUBJECT TO CHANGE AND WITHDRAWAL WITHOUT NOTICE.
- USE OF THIS DOCUMENT IS LIMITED TO SELECT IBM PERSONNEL THIS DOCUMENT SHOULD NOT BE GIVEN TO A CUSTOMER EITHER IN HARDCOPY OR ELECTRONIC FORMAT.

Important notes:

- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.
- IBM makes no warranties, express or implied, regarding non-IBM products and services, including but not limited to Year 2000 readiness and any implied warranties of merchantability and fitness for a particular purpose. IBM makes no representations or warranties with respect to non-IBM products. Warranty, service and support for non-IBM products is provided directly to you by the third party, not IBM.
- All part numbers referenced in this publication are product part numbers and not service part numbers. Other part numbers in addition to those listed in this document may be required to support a specific device or function.
- MHz / GHz only measures microprocessor internal clock speed; many factors may affect application performance. When referring to storage capacity, GB stands for one billion bytes; accessible capacity may be less. Maximum internal hard disk drive capacities assume the replacement of any standard hard disk drives and the population of all hard disk drive bays with the largest currently supported drives available from IBM.

IBM Information and Trademarks

- The following terms are trademarks or registered trademarks of the IBM Corporation in the United States and / or other countries: IBM, IBM Spectrum Storage, IBM Spectrum Protect, IBM Spectrum Scale, IBM Spectrum Accelerate, IBM Spectrum Virtualize, IBM Spectrum Control, Tivoli, IBM Elastic Storage
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Microsoft Windows and the Microsoft Windows Logo are a trademarks or registered trademark of Microsoft Corporation and other countries
- The Redhat logo is a registered trademark of Redhat Inc. in the United States and other countries
- Isilon is a registered trade mark of EMC Corporation in the United States and other countries
- Other company, product, and service names may be trademarks or service marks of others.