Australian Government
Bureau of Meteorology

Bureau of Meteorology –

MARS &
Migration to Linux virtual cluster

Damian Agius – Scientific Computing Services
Commonwealth Bureau of Meteorology

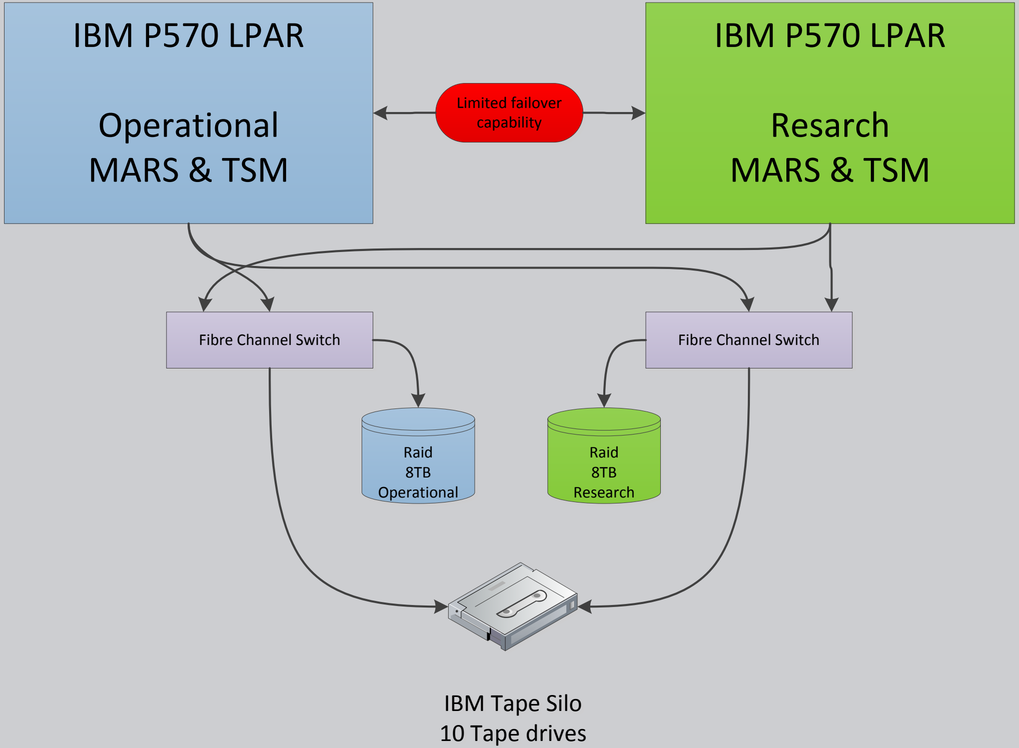
Date: 7th March 2016

Australian Government
Bureau of Meteorology

History of MARS at the Bureau of Meteorology

- 1997: agreement reached with ECMWF to provide MARS software to the Bureau
- 1998: prototyping on IBM RS6000
- 2000-2004: full implementation on IBM SP2 for research department
- 2004-2010: semi-operational on IBM P690
- 2010-Oct 2014: fully operational on IBM P570
- Oct 2014 – Present: MARS operational on virtual machine cluster

Legacy MARS & TSM server



- MARS and TSM on 2x LPAR running on p570
- TSM – 5 x T10000KB drives for operations
- Limited failover capability to run Operational MARS on Research LPAR
- P570 out of support in 2015

MARS Dependencies

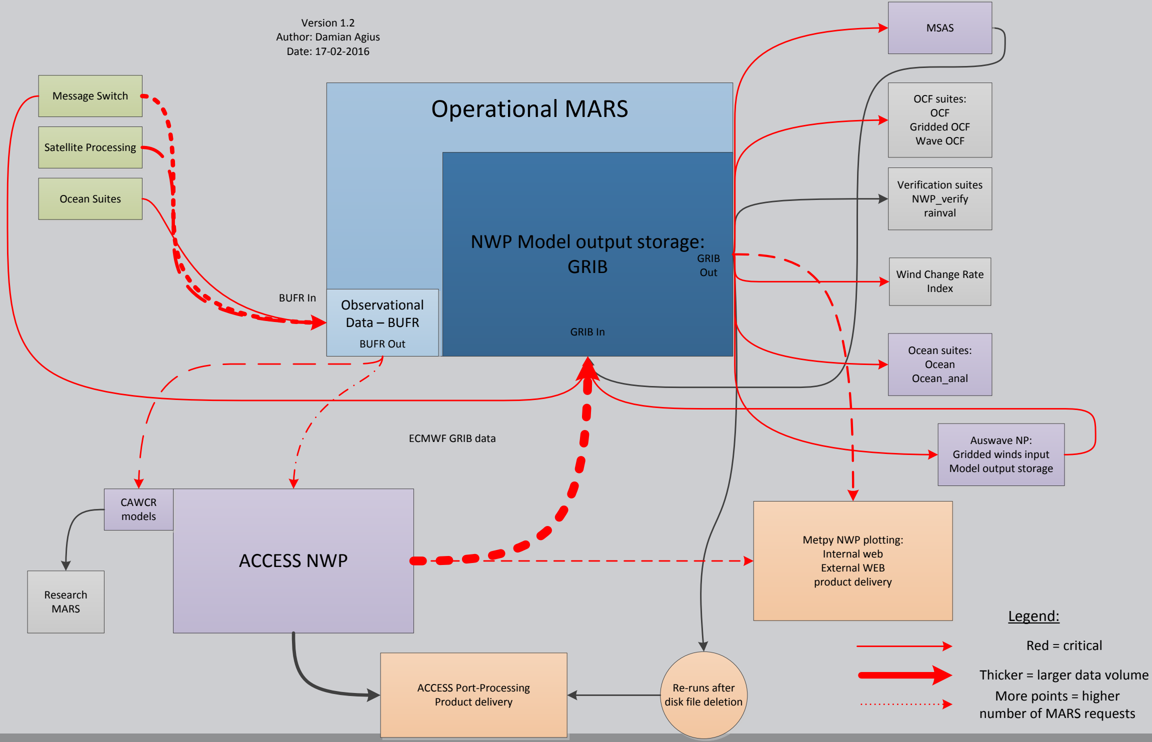
- Majority of operational suites utilise MARS, with some exceptions

- ACCESS NWP output used as input for several other models
  - Majority via MARS
  - Some off disk (non-FDB)

- MARS is in critical path of NWP operations

Operational MARS – Dependencies

Version 1.2
Author: Damian Agius
Date: 17-02-2016

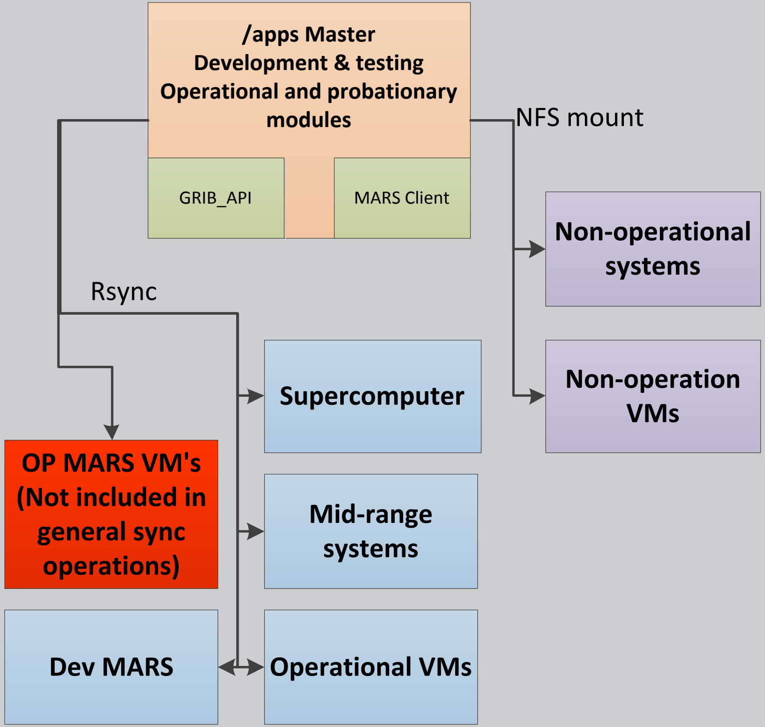


Message Switch
Satellite Processing
Ocean Suites
Operational MARS
NWP Model output storage: GRIB
GRIB Out
BUFR In
Observational Data – BUFR
BUFR Out
GRIB In
MSAS
OCF suites: OCF Gridded OCF Wave OCF
Verification suites NWP_verify rainval
Wind Change Rate Index
Ocean suites: Ocean Ocean_anal
Auswave NP: Gridded winds input Model output storage
ECMWF GRIB data
CAWCR models
Research MARS
ACCESS NWP
Metpy NWP plotting: Internal web External WEB product delivery
ACCESS Port-Processing Product delivery
Re-runs after disk file deletion
Legend:
Red = critical
Thicker = larger data volume
More points = higher number of MARS requests

Mars Client

- Maintained in a Modules environment along with many other apps / libraries / compilers
- Changes to MARS client can be synced to all supported hosts quickly

- Multiple O/S supported:
  - RHEL5
  - RHEL6
  - SLES 11
- NEONS support added to client
  - Gridded -> GRIB & llt data -> BUFR
  - Explained later…

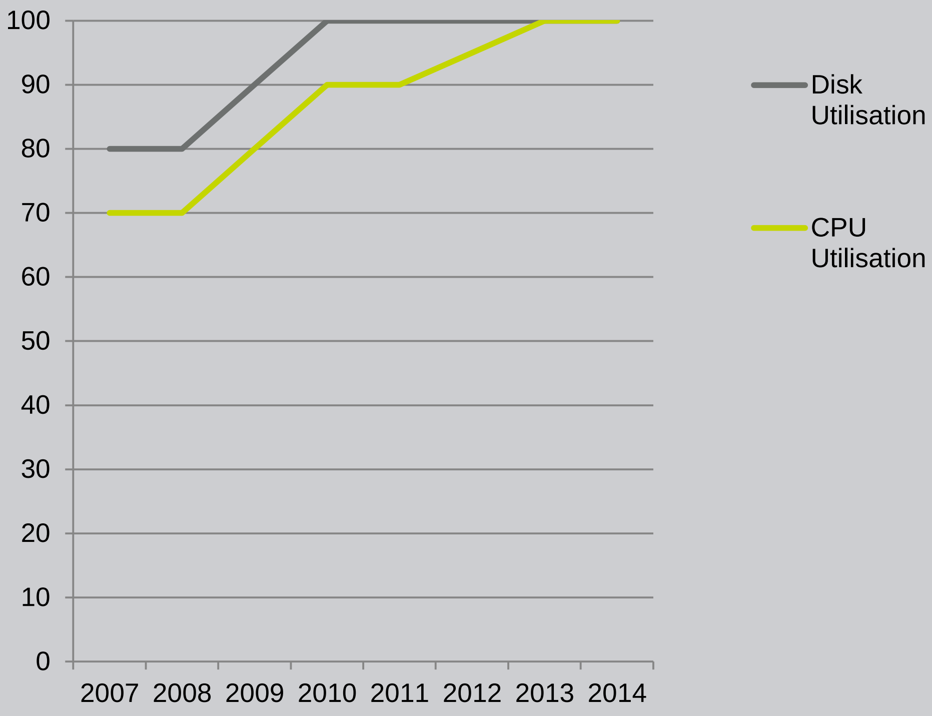


MARS – Migration to Linux Virtual Machine cluster

- Enterprise implementing virtual computing environment
- Testing showed significantly improved MARS performance
- Improved systems management:
  - ability to easily migrate / spin up new MARS nodes in a multiple data centre computing environment

Why The Bureau needed to update MARS

- MARS server version remained at early release
- Regular issues experienced and frequent manual intervention required to maintain operations
- IBM host platform & O/S hitting performance limits
  - Disk
  - CPU
  - IBM p570 support ending 2015
- Only capable of GRIB edition 1 (emoslib / gribex)
- Enterprise moving towards Virtual infrastructure



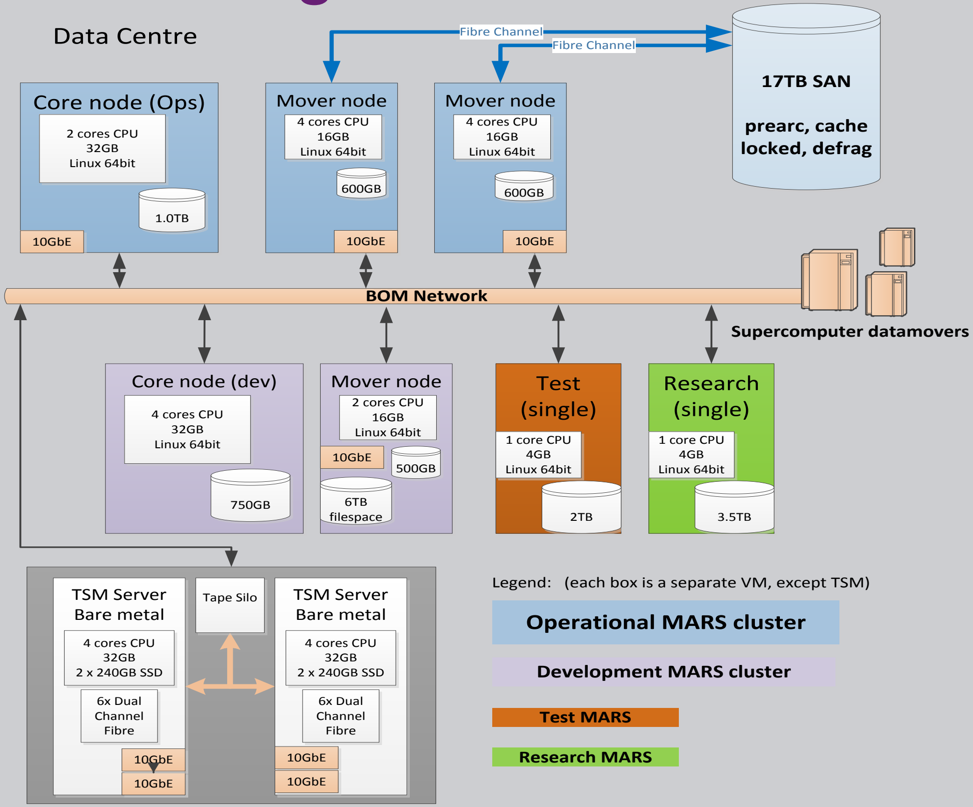
MARS & TSM server configuration

- Old MARS system: Op & Research, with:
  – MARS and TSM on 2x LPAR running on p570

- TSM – 5 x T10000KB drives for operations

- New MARS clusters: Op, dev, test & research
  – Linux Virtual Machines (7 VM servers)
  – 17TB Disk (operational), 8TB dev / test

- Planned Linux based TSM requires bare metal severs co-located with tape Silo.



Data Centre

Core node (Ops)
2 cores CPU
32GB
Linux 64bit
1.0TB
10GbE

Mover node
4 cores CPU
16GB
Linux 64bit
600GB
10GbE

Mover node
4 cores CPU
16GB
Linux 64bit
600GB
10GbE

Fibre Channel

Fibre Channel

17TB SAN

prearc, cache locked, defrag

BOM Network

Supercomputer datamovers

Core node (dev)
4 cores CPU
32GB
Linux 64bit
750GB

Mover node
2 cores CPU
16GB
Linux 64bit
10GbE
500GB
6TB filespace

Test (single)
1 core CPU
4GB
Linux 64bit
2TB

Research (single)
1 core CPU
4GB
Linux 64bit
3.5TB

TSM Server Bare metal
4 cores CPU
32GB
2 x 240GB SSD
6x Dual Channel Fibre
10GbE
10GbE

Tape Silo

TSM Server Bare metal
4 cores CPU
32GB
2 x 240GB SSD
6x Dual Channel Fibre
10GbE
10GbE

Legend: (each box is a separate VM, except TSM)

Operational MARS cluster

Development MARS cluster

Test MARS

Research MARS

Migration considerations

- Subset of downstream NWP operations at BoM are MARS dependant
  - Unable to stop NWP operations for duration of migration activities
  - Needed to be able to extract satellite obs. for input to NWP
  - Needed to use MARS for input to other models (WAVE etc.) and general MetPy plotting
- No changes to operational scripts!
- Flexibility required to assist resolving any unplanned issues discovered during migration activities

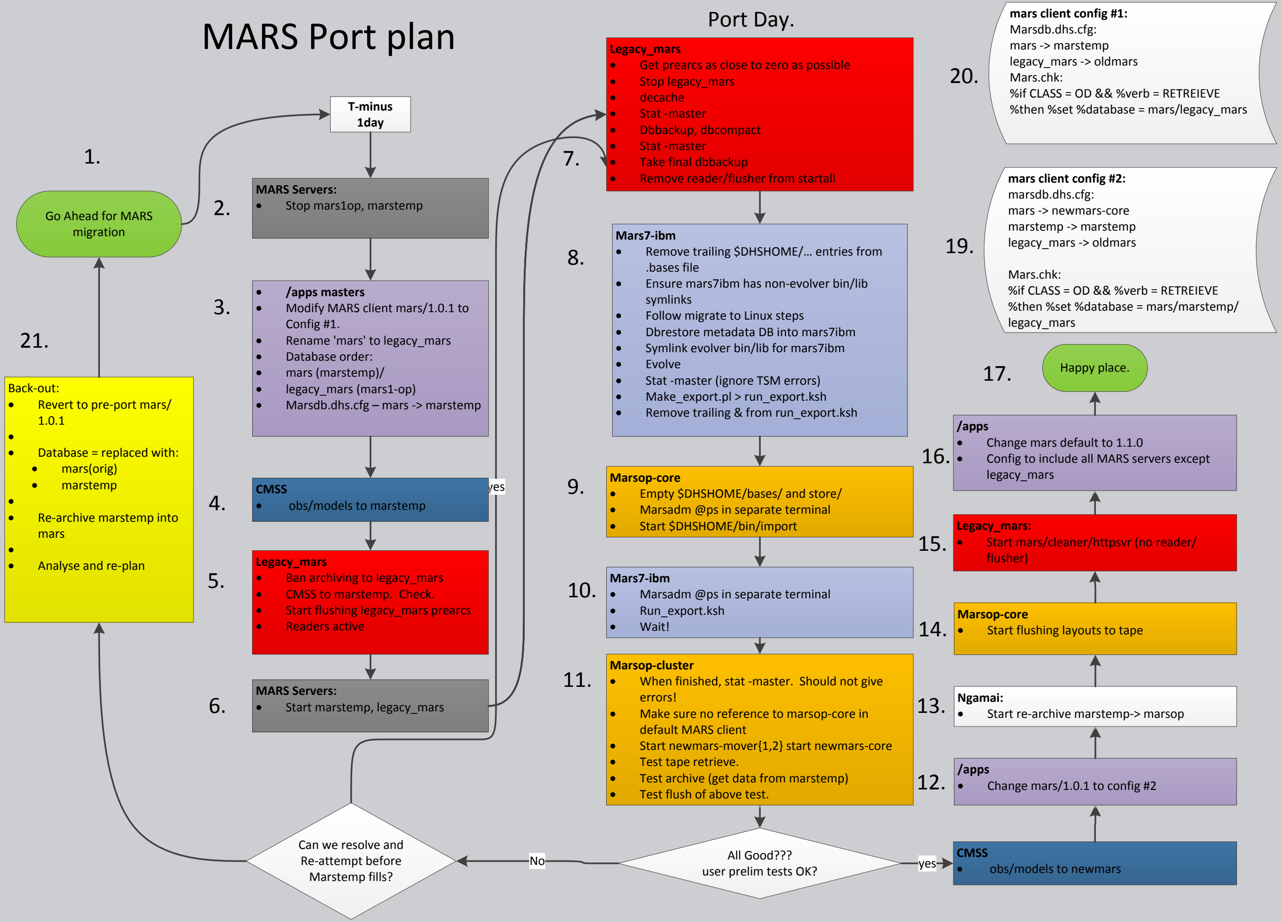
Mitigation techniques

- Create temporary operational MARS server 'marstemp'
  - 10TB disk (5 days capacity)
  - No TSM
- MARS client used to transparently switch between MARS databases
- Used Module environment to sync changes to MARS client across all supported servers as required
  - No modifications to operational scripts
- No changes to total number of operational fields archived during migration activities.

Migration Plan

- Mid Level

MARS Port plan



Port Day.

1. Go Ahead for MARS migration

T-minus 1day

2. MARS Servers:
- Stop mars1op, marstemp

3\.
- /apps masters
- Modify MARS client mars/1.0.1 to Config #1.
- Rename 'mars' to legacy_mars
- Database order:
- mars (marstemp)/
- legacy_mars (mars1-op)
- Marsdb.dhs.cfg – mars -> marstemp

4. CMSS
- obs/models to marstemp

yes

5. Legacy_mars
- Ban archiving to legacy_mars
- CMSS to marstemp. Check.
- Start flushing legacy_mars prearcs
- Readers active

6. MARS Servers:
- Start marstemp, legacy_mars

7. Legacy_mars
- Get prearcs as close to zero as possible
- Stop legacy_mars
- decache
- Stat -master
- Dbbackup, dbcompact
- Stat -master
- Take final dbbackup
- Remove reader/flusher from startall

8. Mars7-ibm
- Remove trailing $DHSHOME/… entries from .bases file
- Ensure mars7ibm has non-evolver bin/lib symlinks
- Follow migrate to Linux steps
- Dbrestore metadata DB into mars7ibm
- Symlink evolver bin/lib for mars7ibm
- Evolve
- Stat -master (ignore TSM errors)
- Make_export.pl > run_export.ksh
- Remove trailing & from run_export.ksh

9. Marsop-core
- Empty $DHSHOME/bases/ and store/
- Marsadm @ps in separate terminal
- Start $DHSHOME/bin/import

10. Mars7-ibm
- Marsadm @ps in separate terminal
- Run_export.ksh
- Wait!

11. Marsop-cluster
- When finished, stat -master. Should not give errors!
- Make sure no reference to marsop-core in default MARS client
- Start newmars-mover{1,2} start newmars-core
- Test tape retrieve.
- Test archive (get data from marstemp)
- Test flush of above test.

All Good??? user prelim tests OK?

yes

No

CMSS
- obs/models to newmars

12. /apps
- Change mars/1.0.1 to config #2

13. Ngamai:
- Start re-archive marstemp-> marsop

14. Marsop-core
- Start flushing layouts to tape

15. Legacy_mars:
- Start mars/cleaner/httpsvr (no reader/ flusher)

16. /apps
- Change mars default to 1.1.0
- Config to include all MARS servers except legacy_mars

17. Happy place.

19. mars client config #2:
marsdb.dhs.cfg:
mars -> newmars-core
marstemp -> marstemp
legacy_mars -> oldmars

Mars.chk:
%if CLASS = OD && %verb = RETREIEVE
%then %set %database = mars/marstemp/ legacy_mars

20. mars client config #1:
Marsdb.dhs.cfg:
mars -> marstemp
legacy_mars -> oldmars
Mars.chk:
%if CLASS = OD && %verb = RETREIEVE
%then %set %database = mars/legacy_mars

Can we resolve and Re-attempt before Marstemp fills?

21. Back-out:
- Revert to pre-port mars/ 1.0.1
-
- Database = replaced with:
  - mars(orig)
  - marstemp
-
- Re-archive marstemp into mars
-
- Analyse and re-plan

Client configurations

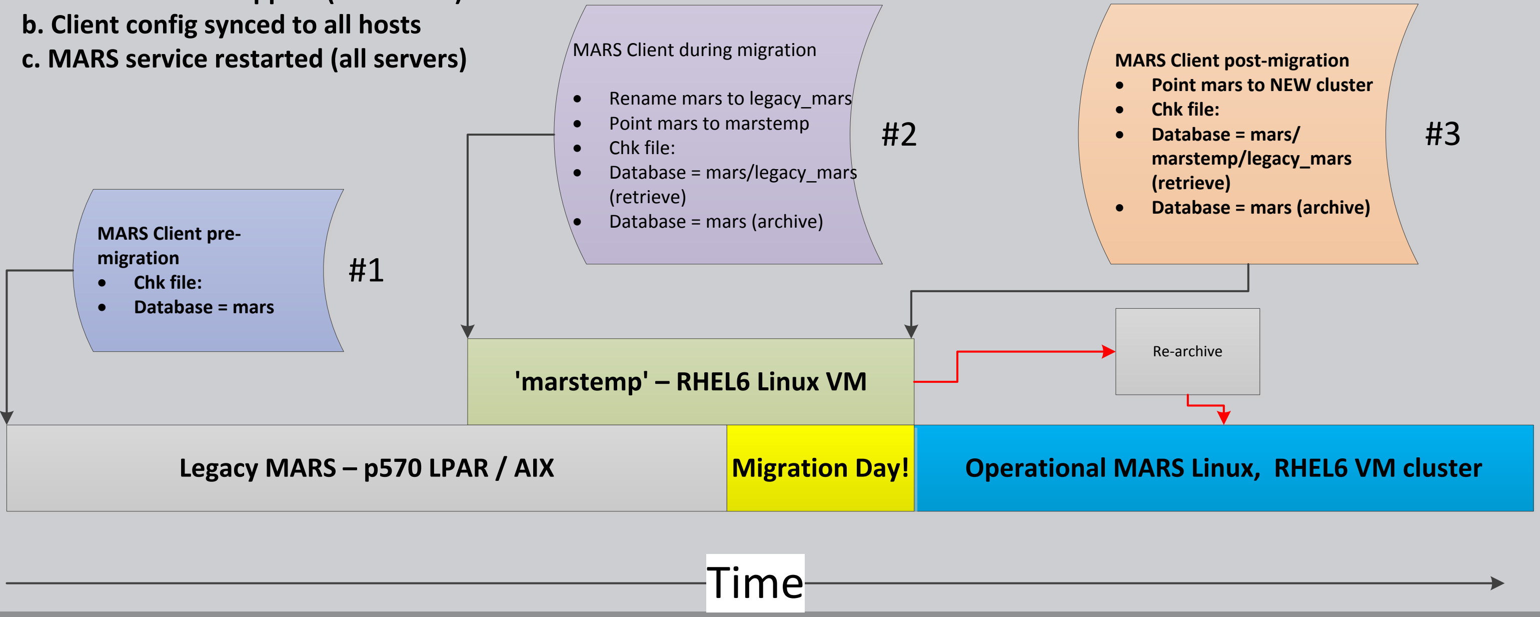
- 3 MARS client configurations used, synced to all MARS client hosts at the appropriate times

- **At each change of MARS client configuration**

a. MARS service stopped (all servers)

b. Client config synced to all hosts

c. MARS service restarted (all servers)



MARS Client during migration

- Rename mars to legacy_mars
- Point mars to marstemp
- Chk file:
- Database = mars/legacy_mars (retrieve)
- Database = mars (archive)

#2

MARS Client post-migration

- **Point mars to NEW cluster**
- **Chk file:**
- **Database = mars/ marstemp/legacy_mars (retrieve)**
- **Database = mars (archive)**

#3

MARS Client pre-migration

- **Chk file:**
- **Database = mars**

#1

'marstemp' – RHEL6 Linux VM

Re-archive

Legacy MARS – p570 LPAR / AIX

Migration Day!

Operational MARS Linux, RHEL6 VM cluster

Time

Testing

- Workflow manager suite for equivalence testing between legacy MARS and new MARS cluster (GRIB) – testing all MARS datwe / layouts
- Scripted testing for BUFR observation retrieval (satellite obs only)
- Modified major NWP suites to use new MARS cluster for testing archive / MetPy plotting / display tasks

- Manual testing other models to ensure retrieval and archiving functioned as expected.
- Load testing I/O between MARS & Super Computer
  - Various prearc file system selection algorithms, 'round robin' fastest during NWP archive periods (less saturation, load spread across mover nodes evenly)

Issues encountered

- MARS client issues:
  - Discovered unexpected format of MARS client requests
  - NEONS interface (for OBS and GRIDS) caused several issues
    - Fixes for both required modification to MARS client configuration 'chk/mars.chk' file and re-sync

- Wave Domains not configured correctly in grib_api
  - For 18 hours, WAVE model used legacy MARS while all other NWP used temporary MARS (using ~/chk/mars.chk)

- Linux Kernel issue encountered, SO_REUSEADDR RHEL6 bug:
- https://access.redhat.com/solutions/357683

  - Issue caused failed flush transactions, particularly for flushes with many files.

    - Some larger layouts were re-archived as few large files, which allowed prearcs to be managed manually (until kernel upgrade performed)

Current use of MARS

- Data Design
- Stats
- Monitoring
- MARS client development, NEONS, GRIB / netCDF integration
- Extending MARS

Data design (Bureau data)

- Class (OD) / Stream
  - One stream per atmospheric model
    - Global
    - Regional
    - City (etc)

- Expver
  - 1,2 = operational
  - 3001/3002 = pre-operational trial
  - 6001/6002 = post-operational

- Type
  - AN,4v : monthly layout

  - FC:
    - Date
    - Levtype
    - Time

BoM MARS archive - operational

- Users:
  - ~ 50 total
    - 6 service accounts
    - 45 normal users (most R&D)

- Daily MARS transactions:
  - 40k - 60k
- Data volumes:
  - 2.0 – 8.0 TB

- 1.5TB Current daily archive volume.
  - 3+TB archive volume forecast by end of 2016

- March 2016 :

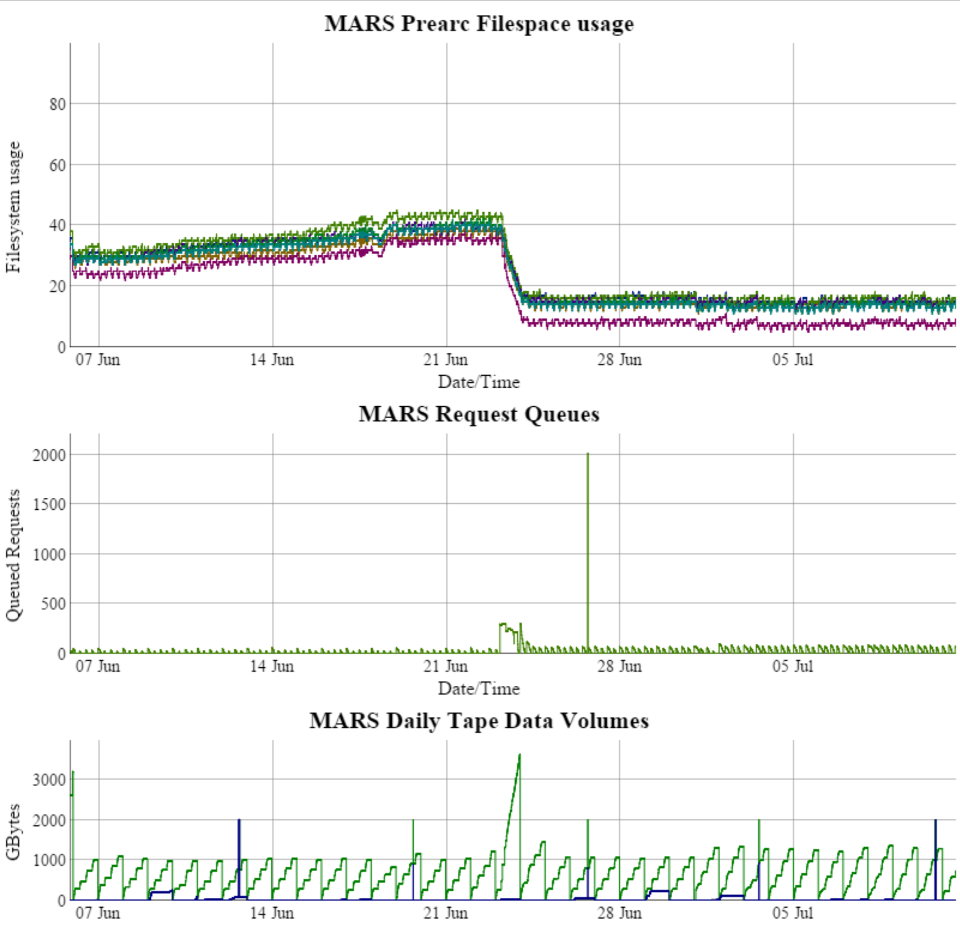
Number of entries	: 444,299
Number of online bytes	: 2.19 Tbytes
Number of offline bytes	: 1.88 Pbytes
Grand total	**: 1.89 Pbytes**
Number of fields	: 3,836,235,947
Number of tape files	: 481,824
Number of disk files	: 112,233
Total number of files	: 594,057
Number of read	: 124,720,842
Oldest read	: 618 weeks

MARS monitoring

- Required metrics collected and plotted using dyGraphs JavaScript package
  - Allows time aligned zooming across plots

- Timestamped marsadm 'ps' and 'df' output placed on web pages

- Operators have controlled and audited access to start/stop MARS



Extending MARS

- The Bureau of Meteorology has extended MARS to integrate other systems into NWP operations
  - MARS client feature additions
  - MARS-like system on R&D supercomputer

- **MARS client:**
  - Support added for extraction from NEONS real time database:
    - NEONS gridded data -> GRIB
    - NEONS llt data -> BUFR
  - Support added to access GRIB files in MARS tree directory structure (similar to FDB)
    - 'database = grbbase'
  - MARS client & openDAP access to NetCDF files in MARS–like directory structure
    - 'database = ncbase'

MARS – like system @ NCI

- Research and Development running on National Computing Infrastructure hosted supercomputer 'Raijin'

- Raijin:
  - Does not have access to tape system with seek ability
  - Does have multi-petabyte disk system

- Research Users have:
  - Requirement for data access and catalog via web such as OpenDAP
  - Requirement to run MARS-coupled apps (MetView, Verify)

Solution:

- MARS server not suitable at NCI, however users still require MARS-like access

- MARS/grib interface developed (similar to the Bureau of Meteorology's MARS / NEONS interface, and FDB)
  - MARS client open/read/write/close feature to any dataset on disk

- GRIB data (and field indexes) stored in directory layout that mirrors the MARS tree i.e.
  - $GRB_ROOT/{op|rd}/<stream>/expver/yyyymmdd/hhhh/an|fc/sfc-fchrs.grb

Australian Government

Bureau of Meteorology

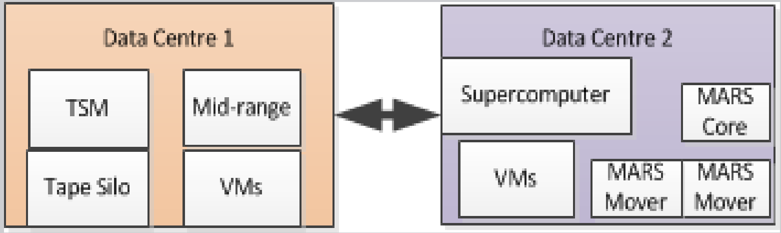
Future of MARS

- Managing MARS VMs across multiple data centres
- Scale MARS for future ACCESS NWP requirements
- Remove MARS from NP critical path

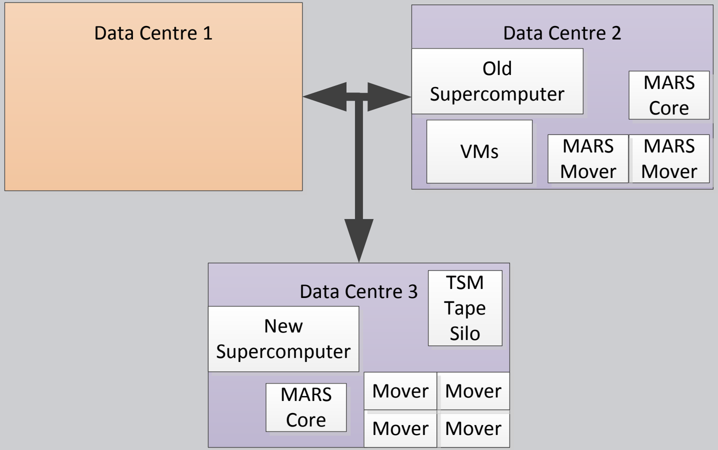
MARS & many data centres

- Currently MARS cluster at one data centre (co-located with current supercomputer), TSM & tape silo at another
  - Ample bandwidth (and network reliability)
  - No issues
- New supercomputer will be at another data centre
  - Parallel trials may require MARS mover nodes at two data centres
  - Final configuration likely redundant operational MARS clusters in two data centres

Current



Future

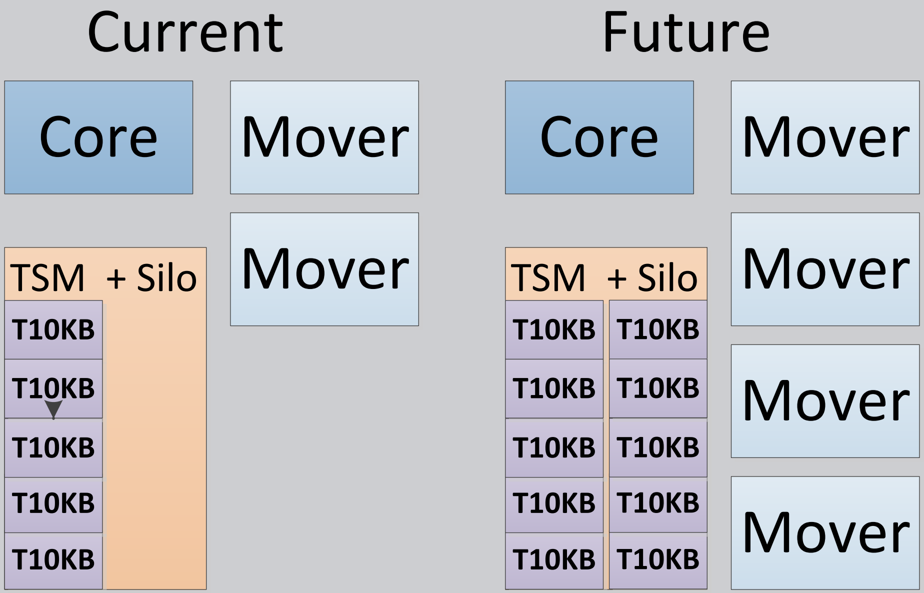


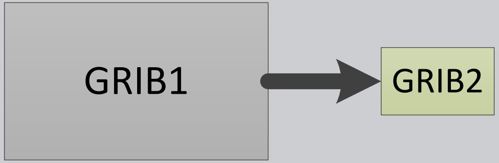
Scaling MARS

- Adding more mover nodes
  - Mover node disk sizing at preferred maximum size for VM management

- Doubling tape drives available to TSM
  - Currently 5 x T10KB
  - Mid-2016 10 x T10KB

- Archiving in GRIB edition 2
  - No real tape savings, however less bandwidth & disk space used during distribution & archiving





Thank you…

- Damian Agius
- +61396694387
- d.agius@bom.gov.au