2015 DMF USER GROUP

DMF WITHOUT TAPES

Is it a good idea to create an HSM without tapes?

ALAN DAVIS

CENTRE FOR BIOIMAGING SCIENCES (CBIS)
MECHANOBIOLOGY INSTITUTE (MBI)
NATIONAL UNIVERSITY OF SINGAPORE

CBIS NUS CENTRE FOR BIOIMAGING SCIENCES

MBI MECHANOBIOLOGY INSTITUTE National University of Singapore

NUS National University of Singapore

Previous Data Requirements and Assumptions

Started life as a scientist – so still try to think like one

Biologists generally have modest IT needs except for

- BioInformatics
- BioImaging

Previously at MIT – managing one BioImaging lab

- 3 x Optical microscopes, 1 x Electron microscope
  - Detectors collected ~ 10-100 MB/s
- 1 PI and 20-30 users
- resulted in - 100 TB / 5 years

Coming to Singapore – managing 2 Research Centres

- 30+ Optical microscopes, 4 x Electron microscopes
  - Detectors can collect 1+ GB/s
- 25 PIs and 300 users
- Planned for 10x growth over MIT lab
- Anticipated 1 PB / 5 years
- SMALL server room, 2-3 racks storage

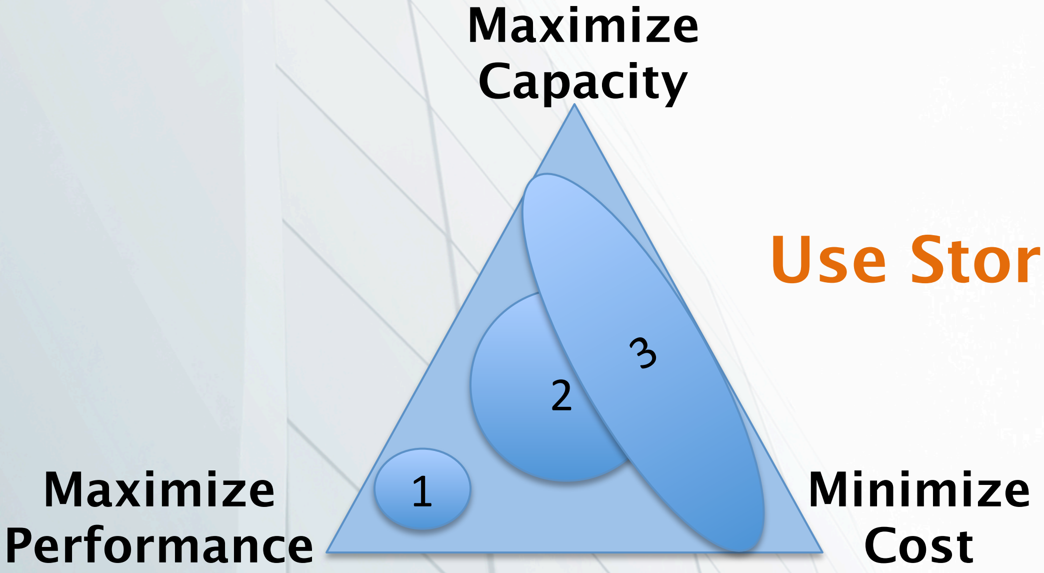
How to design a PB storage system

Basic storage design criteria – Try to Balance the following:

- Capacity
- Performance
- Cost

Gating Factors

- Ease of Management
- Can't lose any user data



Use Storage Tiering

How to design a PB storage system

Non- Tape based Tiered Storage Solution

	Performance	Capacity	Cost
Tier 1	Maximum	Minimum	High
Tier 2	Moderate	Moderate	Moderate
Tier 3	Moderate	High	Low

HSM to manage data movement between the tiers.

How to design a PB storage system

2010-2014 HDD characteristics

Performance

- SAS excellent for types of I/O: sequential, random, large, small
- SATA excellent for sequential, large I/O
- NL-SAS excellent for sequential, large, I/O, good for random , small I/O

Capacity

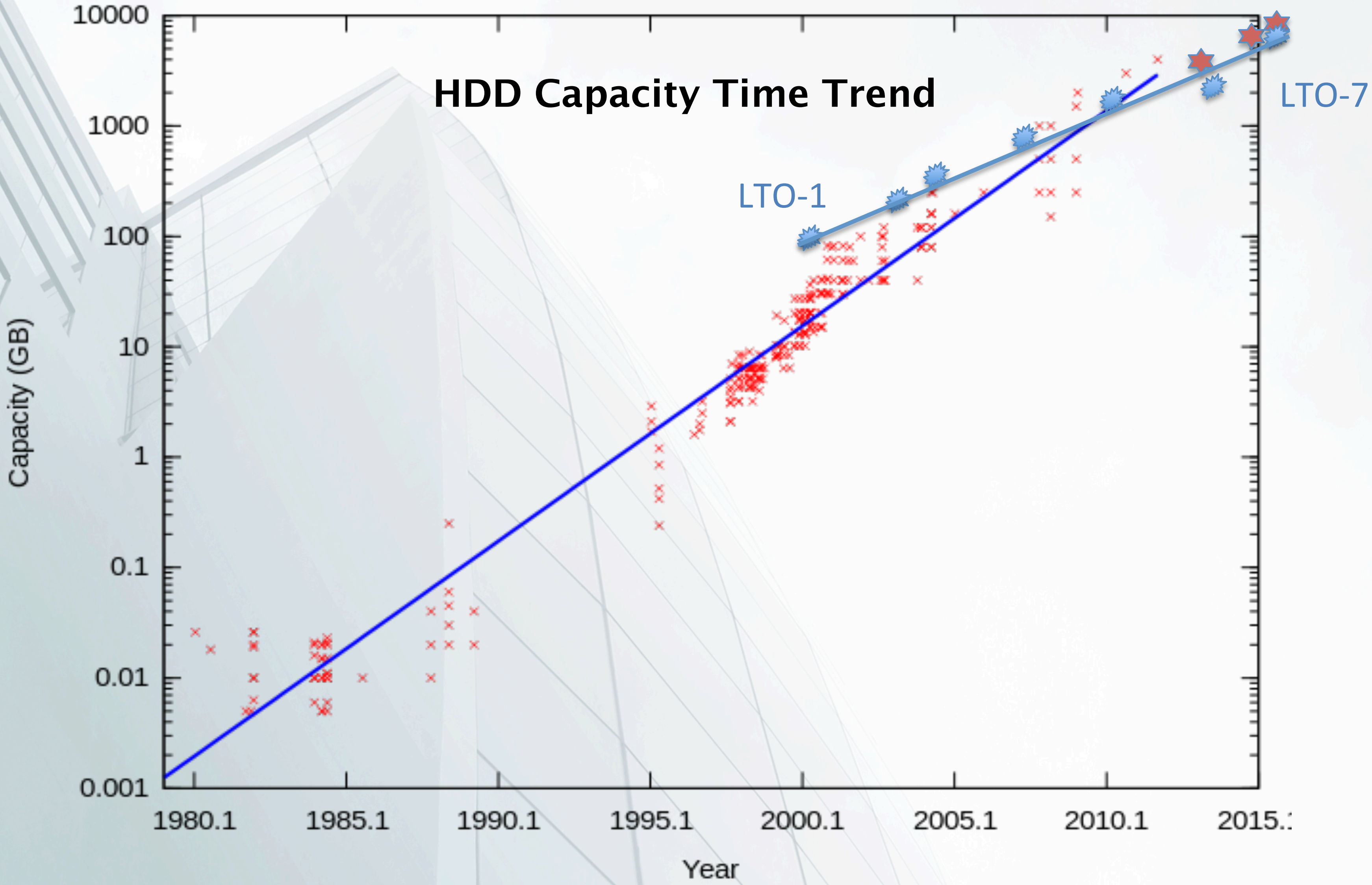
- SATA, NL-SAS – 2, 3 TB

Cost of Tier

- 1- SAS $1000+ / TB
- 2- SATA, NL-SAS - ~ $500-700 / TB
- Est. cost for 500 TB = ~ $500K /lab fit in my 5yr budgets ☺

2014 – 4 TB, Enterprise, NL-SAS < $100/TB

How to design a PB storage system



How to design a PB storage system

DMF Characteristics - 2009

- **Tier 1 – XFS filesystem – needs DMAPI mount option (future ?)**
- **Tier 2 – MSP -> DCM MSP – XFS filesystem needs DMAPI**
- **Tier 3 – FTP / Tape / COPAN MAID / Cloud (201?)**

- **? DMF compatible with NAS/NFS based storage devices**

  - **2012 – SGI-DMF_5.6_AdminGuide – 007-5484-010 p. 11**

    DMF **interoperates** with the following:

    - Standard data export services such as ***Network File System (NFS)*** and FileTransfer Protocol (FTP)
    - XFS® filesystems
    - CXFS™ (clustered XFS) filesystems

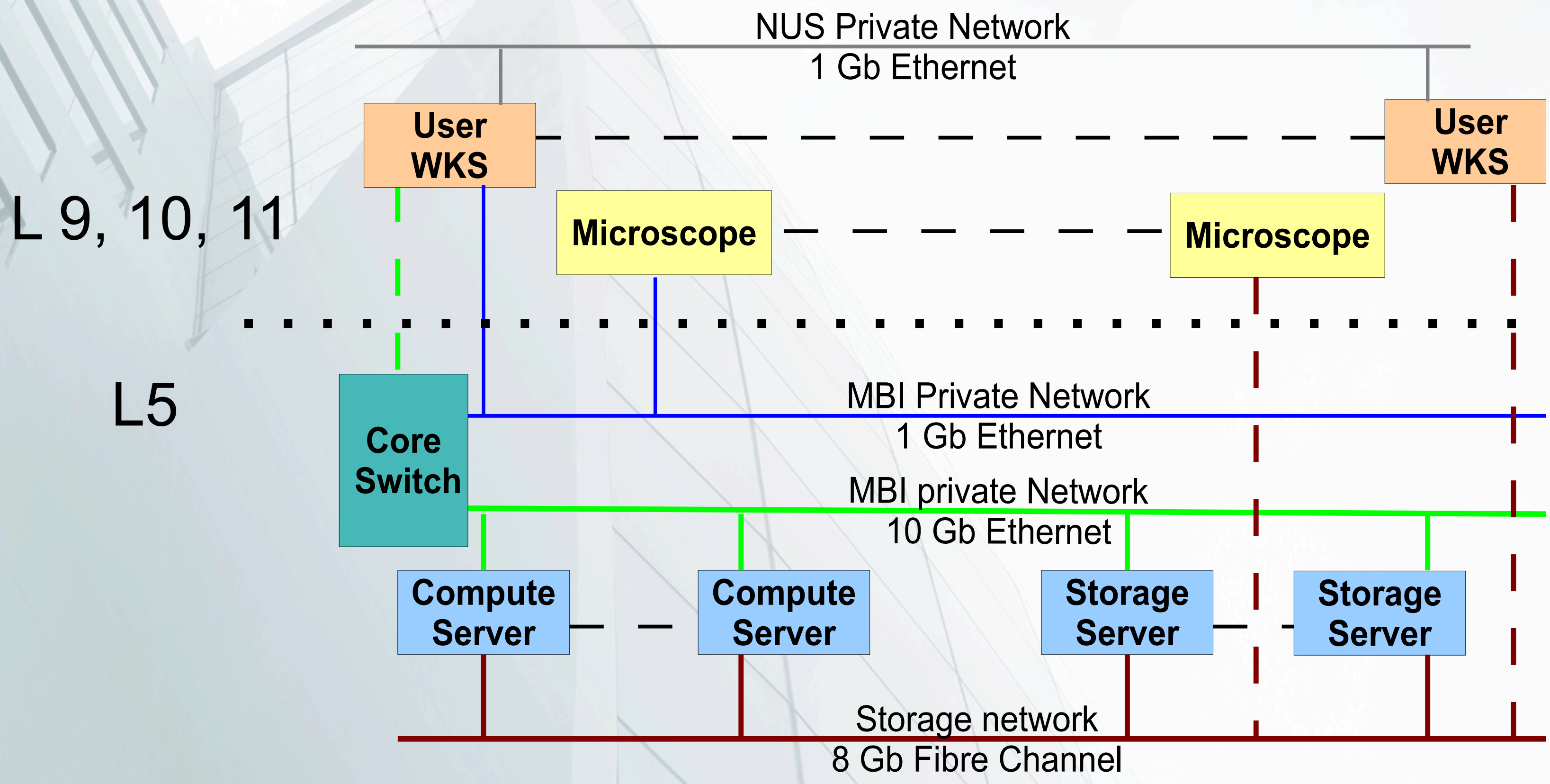
  - NO mention as Tier 3 target
  - Still TRUE in 2015 DMF v6

MBI System layout

1 GbEthernet

10 Gb Ethernet

8 Gb FibreChannel



Storage System Evolution – 2010 to 2014

- **2010** – **MBI**, **1 tier** + NAS **(63-0-24 TB)** using 1x DDN 6600 (2TB SATA)
  - w SSD for XFS inodes
- **2010** – **CBIS**, **2 tiers**, (**31-110-0 TB**) with DMF, using 1x DDN 6600
- **2011** – **MBI** capacity expansion (**123-0-24 TB**) DDN 6600 exp encl
- **2012** – **CBIS** capacity expansion (**28-219-0 TB**), 2nd DDN 6600
- **2013** – **MBI**, **2nd tier** & capacity expansion (**28-214-48 TB**) with NetAPP 5500
- **2014** – **MBI**, **3rd tier** using cluster of storage servers, 4x 64 TB (**256 TB**) using **BeeGFS**
  - eval'd **Ceph FS,** thought about **Lustre,** dismissed **GlusterFS** and **GPFS**
- **2014** – **CBIS**, **3rd tier** with **SGI MIS** box, (**224 TB**)

Storage System Evolution – 2010 to 2014

Total Storage Usage: MBI and CBIS

	Tier 1	Tier 2	Tier 3	Total (TB)
MBI	23	168	160	351
CBIS	24	197	156	377
Total	47 / 55	365 / 433	316 / 480	728 / 968

Storage System Evolution – 2010 to 2014

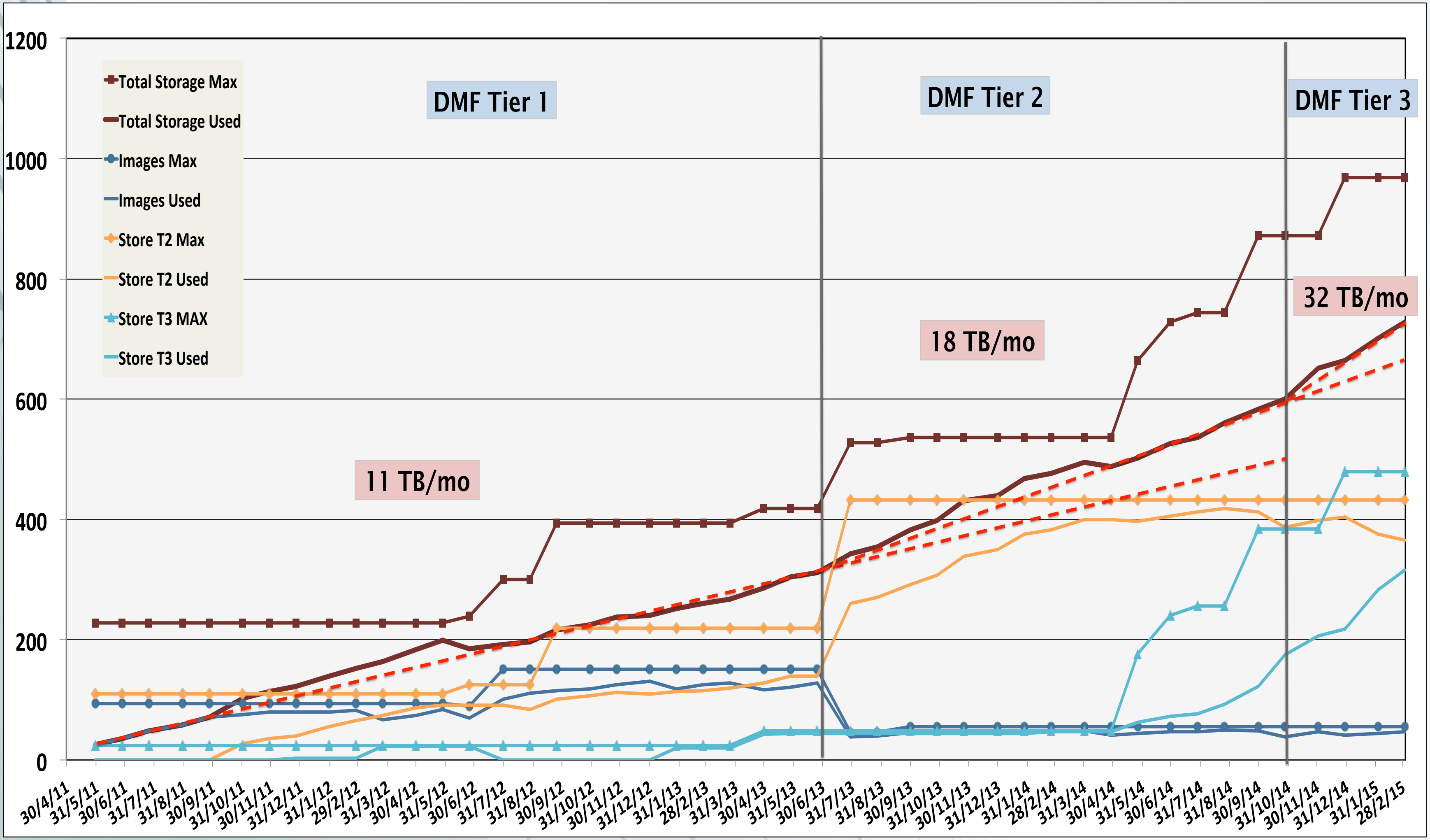
Tier Performance

	Tier 1	Tier 2	Tier 3
	R5 SAS (4+1)	R6 NL-SAS (8+2)	NFS
MBI	1 – 2* GB/s	.5 – 1* GB/s	.2 - .3 GB/s 15 TB/d
CBIS	1 - 2 GB/s	.5 - 1 GB/s	.2 - .5 GB/s 20 TB/d

* Upgrading to DDN 7700 = 6 - 8 GB/s

Storage System Usage

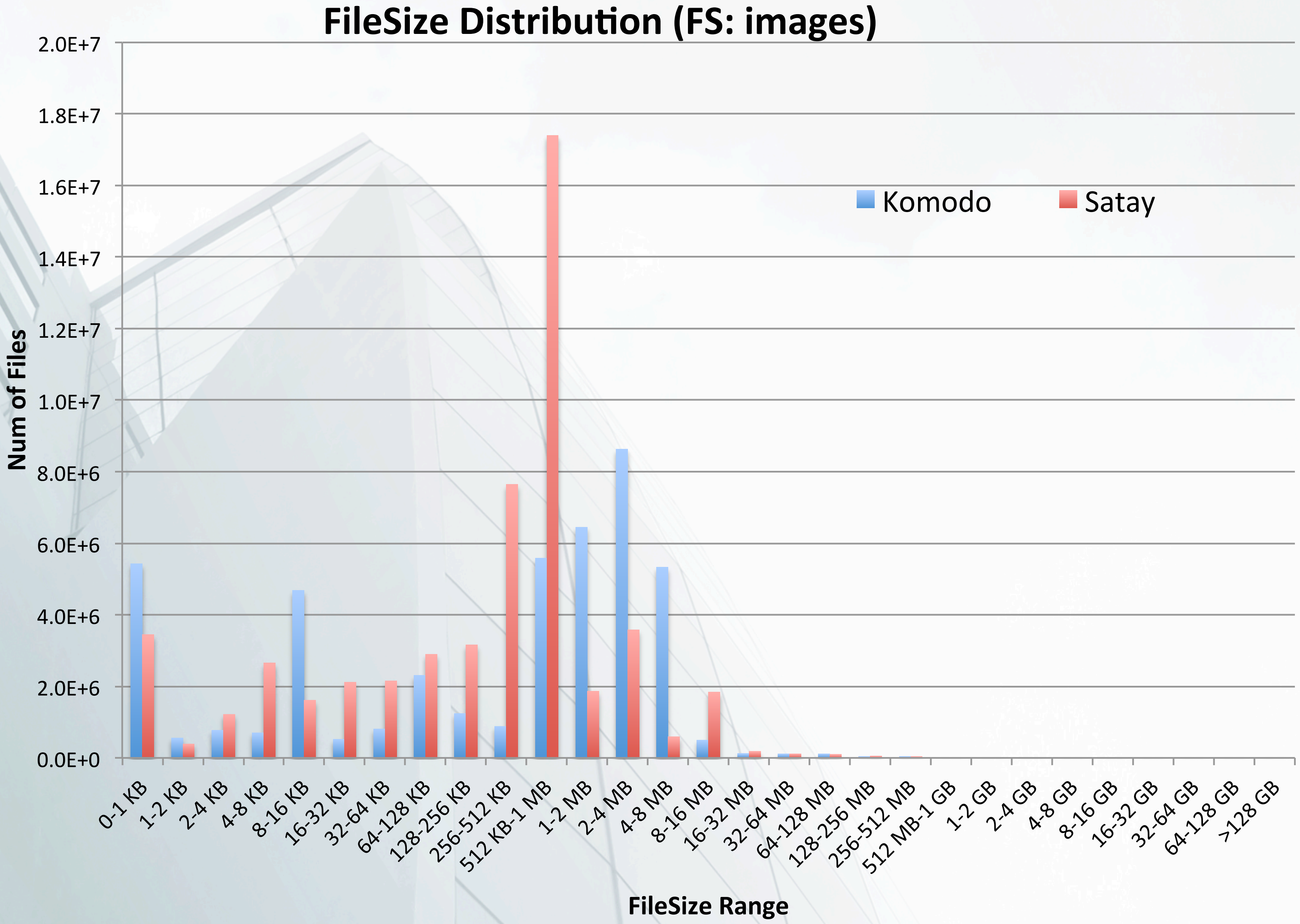
MBI/CBIS Storage Usage - 3 tiers: Images (28+28 TB), Tier2 (214+219 TB), Tier3 (256+224 TB



MBI Data Increase (TB / Mon)

CBIS NUS CENTRE FOR BIOIMAGING SCIENCES

MBI MECHANOBIOLOGY INSTITUTE National University of Singapore


FileSize Distribution (FS: images)
Num of Files
2.0E+7
1.8E+7
1.6E+7
1.4E+7
1.2E+7
1.0E+7
8.0E+6
6.0E+6
4.0E+6
2.0E+6
0.0E+0
Komodo
Satay
0-1 KB
1-2 KB
2-4 KB
4-8 KB
8-16 KB
16-32 KB
32-64 KB
64-128 KB
128-256 KB
256-512 KB
512 KB-1 MB
1-2 MB
2-4 MB
4-8 MB
8-16 MB
16-32 MB
32-64 MB
64-128 MB
128-256 MB
256-512 MB
512 MB-1 GB
1-2 GB
2-4 GB
4-8 GB
8-16 GB
16-32 GB
32-64 GB
64-128 GB
>128 GB
FileSize Range

Current DMF configuration

with thanks from Susheel Gokahle

I didn’t lose any data during this process ☺

Based on DMF sample file dmf.conf.dcm

```
define      daemon
            TYPE                            dmdaemon
            MIGRATION_LEVEL                 auto
#       The following parameter must not be changed while DMF is running!
            MSP_NAMES                       dsk1 dsk1_t3L dsk1_t3A          * dmcheck complains
            TASK_GROUPS                     daemon_tasks dump_tasks
            MOVE_FS                         /dmf/spare
            EXPORT_QUEUE                    ON
            MESSAGE_LEVEL                   2
#       Turn off partial access to files being recalled when used with CXFS, see DMF Admin Guide
            RECALL_NOTIFICATION_RATE                  0
enddef
```

Current DMF configuration – Tier 1

```
define   /mnt/mbi/images
         TYPE                    filesystem
         MIGRATION_LEVEL         auto
         POLICIES                space_policy msp_policy
         USE_UNIFIED_BUFFER      OFF
         BUFFERED_IO_SIZE        1048576
         DUMP_MIGRATE_FIRST      OFF
         MESSAGE_LEVEL           2
enddef

define   space_policy
         TYPE                    policy
         FREE_SPACE_MINIMUM      10
         FREE_SPACE_TARGET       20
         MIGRATION_TARGET        95
         FREE_DUALSTATE_FIRST    OFF
         AGE_WEIGHT              -1      0          when uid in (root mbiftp mbitest)
         AGE_WEIGHT               1      1
         SPACE_WEIGHT            -1      0          when size <= 65536
         SPACE_WEIGHT             0      .00000001
enddef
```

Current DMF configuration – Tier 2 DCM

```
#
#          Define the "dcm_msp" as a disk MSP.
#          Keep the same name "dsk1" as before in the 2-tier config
#
define     dsk1
           TYPE                          msp
           COMMAND                       dmdskmsp
           MIGRATION_LEVEL               auto
           POLICIES                      dcm_space_policy
           TASK_GROUPS                   dcm_tasks
           STORE_DIRECTORY               /dmf/store
           BUFFERED_IO_SIZE              1048576
           CHILD_MAXIMUM                 12
           GUARANTEED_GETS                4
           NAME_FORMAT                   %u/%y/%m/%d/%b
           MESSAGE_LEVEL                 2
           DUMP_FLUSH_DCM_FIRST          OFF
           WRITE_CHECKSUM                ON
enddef
```

Current DMF configuration – Tier 2 DCM

```
#
#        Define how the cache will be managed, writing 2 copies to archive storage.
#
define   dcm_space_policy
         TYPE                       policy
         FREE_SPACE_MINIMUM          5
         FREE_SPACE_TARGET          10
         DUALRESIDENCE_TARGET       40
         FREE_DUALRESIDENT_FIRST    ON
         CACHE_AGE_WEIGHT           -1        0             when age < 180
         CACHE_AGE_WEIGHT            1        1
         CACHE_SPACE_WEIGHT         -1        0             when size <= 262144
         CACHE_SPACE_WEIGHT         0         .000000001
         SELECT_LOWER_VG            none                    when softdeleted = true
         SELECT_LOWER_VG            dsk1_mgL                when size < 1         *dmcheck fix
         SELECT_LOWER_VG            dsk1_mgR                                      *dmcheck fix
enddef
```

Current DMF configuration – Tier 3 NFS

```
#
#         Define a migrate group that is comprised of the 2 DISK MSPs
#
define    dsk1_mgL
          TYPE                        migrategroup
          GROUP_MEMBERS               dsk1_t3L
          ROTATION_STRATEGY           SEQUENTIAL
enddef

#
#         For duplicate copies at remote site directory on SGI MIS at CBIS
#
define    dsk1_mgR
          TYPE                        migrategroup
          GROUP_MEMBERS               dsk1_t3A
          ROTATION_STRATEGY           SEQUENTIAL
enddef
```

Current DMF configuration – Tier 3 NFS

```
#
#       Define the dsk1_tier3 msps.  You must modify the STORE_DIRECTORY parameter to
#       a value appropriate for your site.  The remote sites are NFS mounted directories
#
define  dsk1_t3L
        TYPE                    msp
        COMMAND                 dmdskmsp
        STORE_DIRECTORY         /dmf/store_tier3L
        FULL_THRESHOLD_BYTES    255750000000000
        CHILD_MAXIMUM           8
        NAME_FORMAT             %u/%y/%m/%d/%b
        MESSAGE_LEVEL           2
        WRITE_CHECKSUM          ON
enddef
```

Current DMF configuration – Tier 3 NFS

```
#
#       For duplicate copies at remote site directory on SGI MIS at CBIS
#
define  dsk1_t3A
        TYPE                     msp
        COMMAND                  dmdskmsp
        STORE_DIRECTORY          /dmf/store_tier3A
        FULL_THRESHOLD_BYTES     223900000000000
        CHILD_MAXIMUM            8
        NAME_FORMAT              %u/%y/%m/%d/%b
        MESSAGE_LEVEL            2
        WRITE_CHECKSUM           ON
enddef
```

Current DMF configuration - dmcheck

```
Checking DMF installation.
  Linux satay 3.0.101-0.35-default #1 SMP Wed Jul 9 11:43:04 UTC 2014 (c36987d) x86_64
x86_64 x86_64 GNU/Linux - satay
  SuSE-release:          SUSE Linux Enterprise Server 11 (x86_64)
  SuSE-release:          VERSION = 11
  SuSE-release:          PATCHLEVEL = 3
  sgi-issp-release:      SGI InfiniteStorage Software Platform, version 3.2, Build
710r1.sles11sp3-1407021914
  sgi-foundation-release:         SGI Foundation Software 2.10, Build
710r16.sles11sp3-1404092103
  lsb-release:           LSB_VERSION="core-2.0-noarch:core-3.2-noarch:core-4.0-noarch:core-2.0-
x86_64:core-3.2-x86_64:core-4.0-x86_64"
  DMF version 6.2.0 rpm dmf-6.2.0-sgi320rp32.sles11sp3 installed.

Checking DMF config file /etc/dmf/dmf.conf

 Scanning for non-comment lines outside define/enddef pairs
 Scanning for DMF parameters without values
 Checking all objects for invalid names
 Checking base
 Checking DMF license
  Total bytes managed            505105182990336 (505TB)
  Total charged to license       349955270135890 (349TB)
  DMF license capacity          500000000000000 (500TB)
  Percent of license capacity              69
```

Current DMF configuration - dmcheck

```
Checking daemon
Checking policy dcm_space_policy
Checking policy msp_policy
 WARNING:  Some files will only have one copy made when migrated.
Checking policy space_policy
Checking filesystem /mnt/mbi/images
Checking filesystem /dmf/archive
Checking MSP dsk1 (DCM-mode)
 WARNING:  dsk1 follows LS dsk1_t3L (containing VG dsk1_t3L) on the
              LS_NAMES/MSP_NAMES parameter; recalls will bypass it
Checking MSP dsk1_t3L
Checking MSP dsk1_t3A
Checking Migrate Group dsk1_mgL
Checking Migrate Group dsk1_mgR
Checking selection rules in policy space_policy.
Checking selection rules in policy msp_policy.
Checking selection rules in policy dcm_space_policy.
```

* dmcheck issue ?

Current DMF configuration - dmcheck

```
 Checking Services services
 Checking Task Group daemon_tasks
 Checking Task Group dump_tasks
 Checking Task Group dcm_tasks
 Checking for unreferenced objects

 Checking other daemons.
 Checking chkconfig
  WARNING:  dmf is disabled by chkconfig

No Errors found.
3 warnings found.
```

Note: fixed a dmcheck bug – unable to find xvm local volumes (eg /dev/lxvm/dmfstore) when CXFS is running

DMF *DOESN'T NOT* complain

Future Storage Growth

2010 thru 2014 – Resulted in 728 TBs

for 2015-2019 projecting: 300-400 TBs / yr

Do the original assumptions still hold?

Can continue with HDD based tier 3s to: 2, 3, 5, … PBs ?

Storage vendors project HDDs to 50+ TB in a few years!

Do we want to use drives that BIG?

- Need more intelligent drive rebuilds
- Standard Raid 5 or 6 is inadequate
- Raid pools for rebuilds
- How will that affect tuning and I/O optimizations

Future Storage Growth

Future of storage drives, systems and software

Convergence of Storage Servers and Storage Arrays

- Need to reduce the movement of data between tiers
- Combine the functionality of storage servers and storage arrays
- Put storage filesystem software onto the storage arrays

HSM tiering – Can we converge some number of tiers (eg 2 & 3)

- Can we use HDDs for archival tiers of capacity: 10, 20, 50 PBs?
- Can it provide the reliability, DR to remote site, restoration times, etc

Singapore is purchasing a 1-3 PFLOP SC with 10-20 PBs of storage

- Storage vendors may be proposing an all HDD based solution

DMF Without Tapes

Was it a good idea – Comments

Questions ??

Al Davis
aldavis@nus.edu.sg