Why HSMs still have a place at ECMWF

Tape is dead! Or is it?

Francis Dequenne
francis.dequenne@ecmwf.int

ECMWF

ECMWF

European Centre for Medium-range Weather Forecast

June 2, 2015

What is ECMWF?

European Centre	We are an independent international organisation funded by 34 States
Medium-Range	Up to fifteen days ahead. Today our products also include monthly and seasonal forecasts and we collect and store meteorological data.
Weather Forecasts	We produce world-wide weather forecasts

European collaboration at its best!

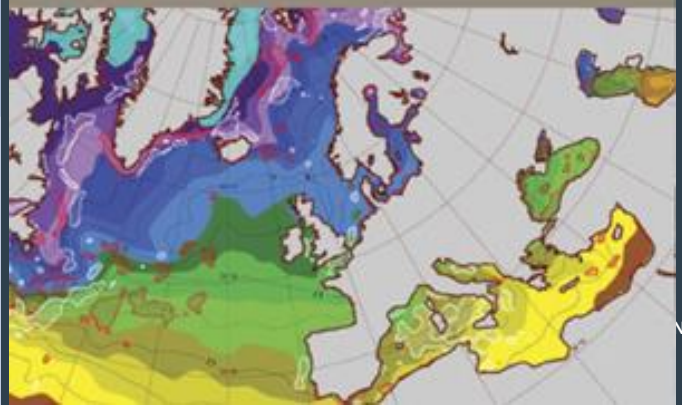
ATMOSPHERIC COMPOSIT

SUPERCOMPUTER CENTRE

CLIMATE MONITORING

- Global numerical weather forecasts
- Composition of the atmosphere: monitoring and forecasting
- Climate reanalysis: monitoring
- Supercomputing & data archiving
- Education programme

GLOBAL PREDICTION



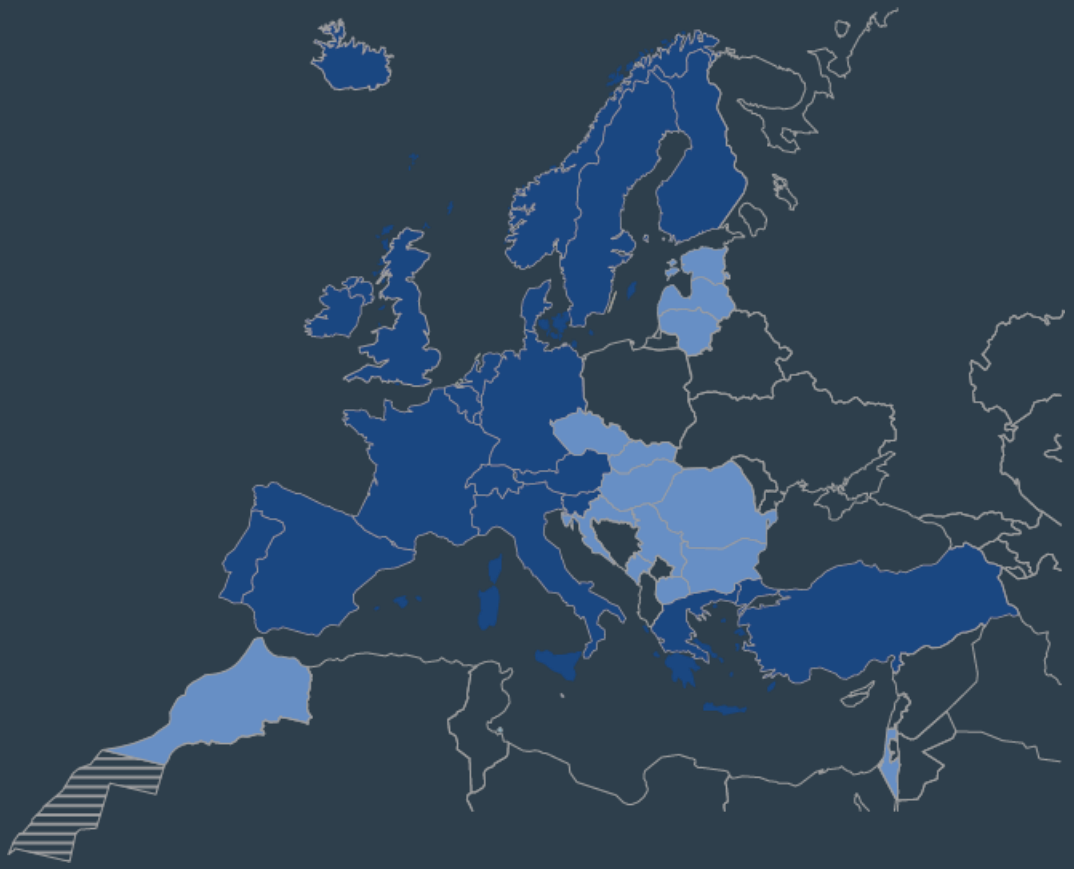
NTRE FOR MEDIUM-RANGE WEATHER FORECASTS

SEVERE WEATHER

European with a global reach

- 34 member and co-operating states
- 270 staff
- 30 countries
- Partnerships around the world …

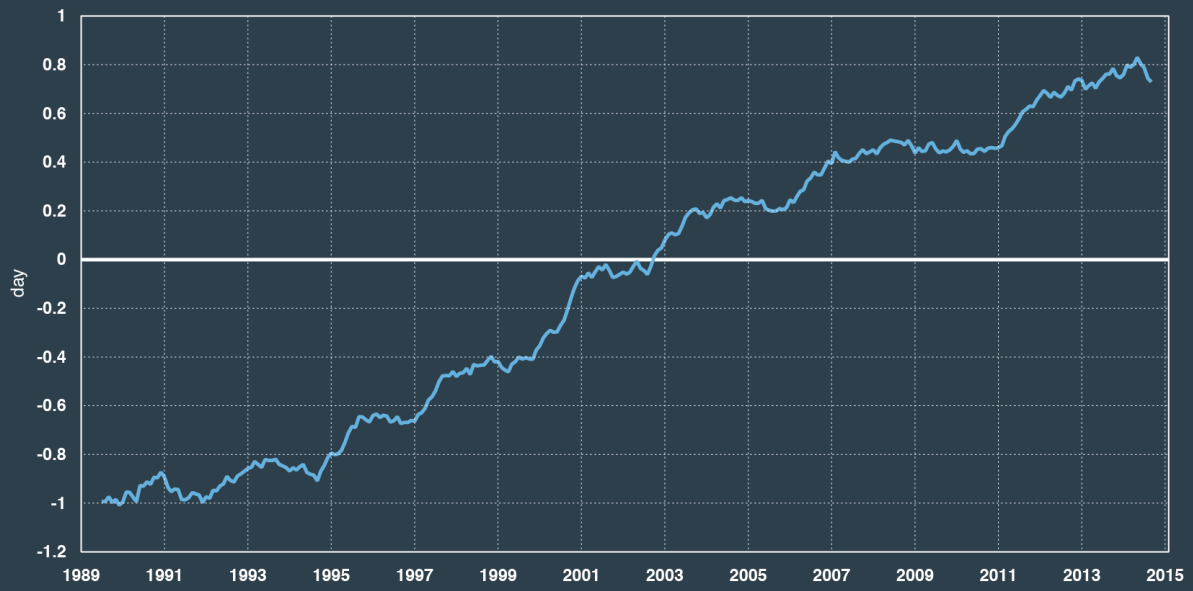
International cooperation at its best



ECMWF

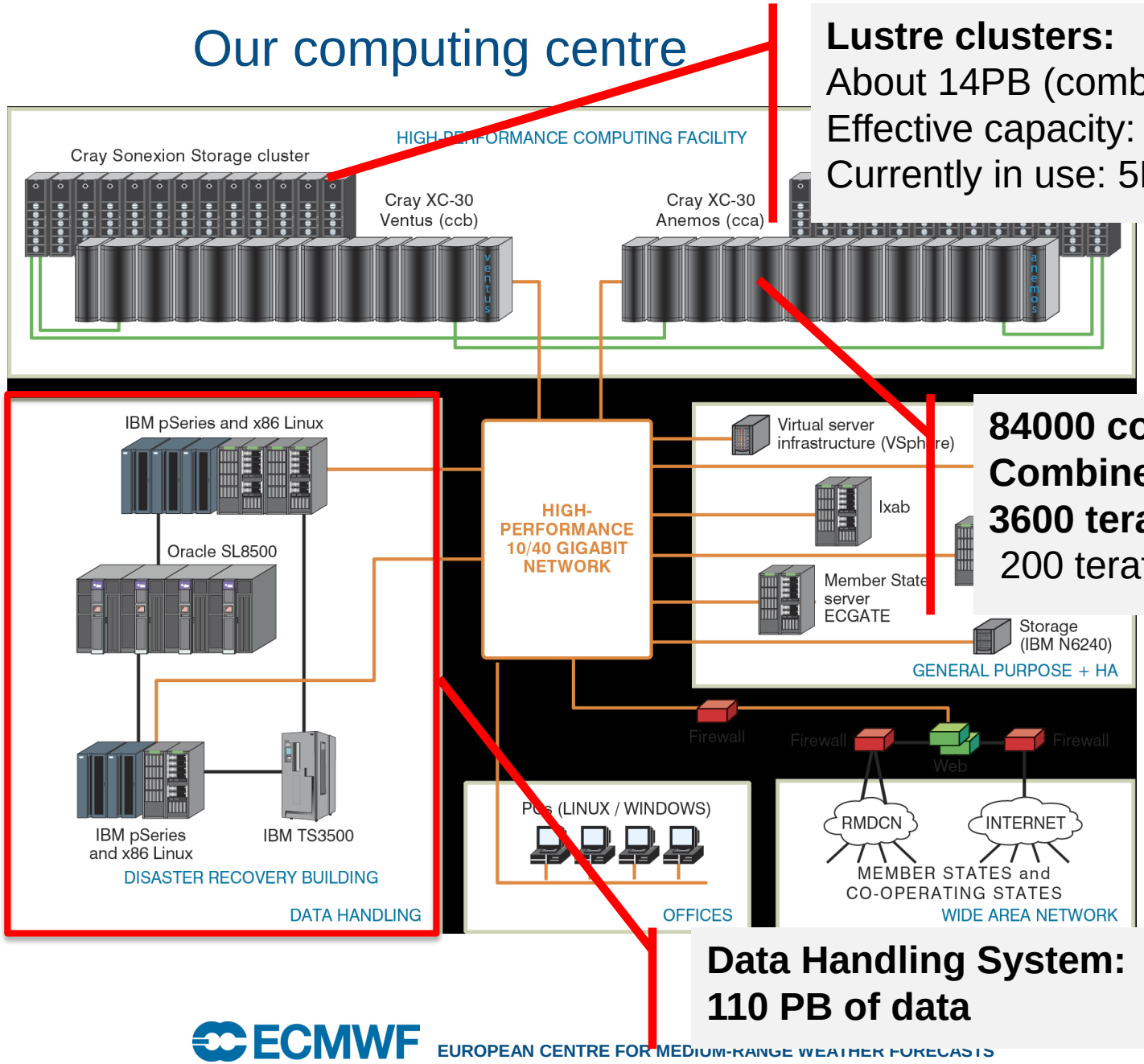
Mission-driven science

Change in the range of skilful forecasts compared to using the operational system of ten years ago



ECMWF

Our computing centre

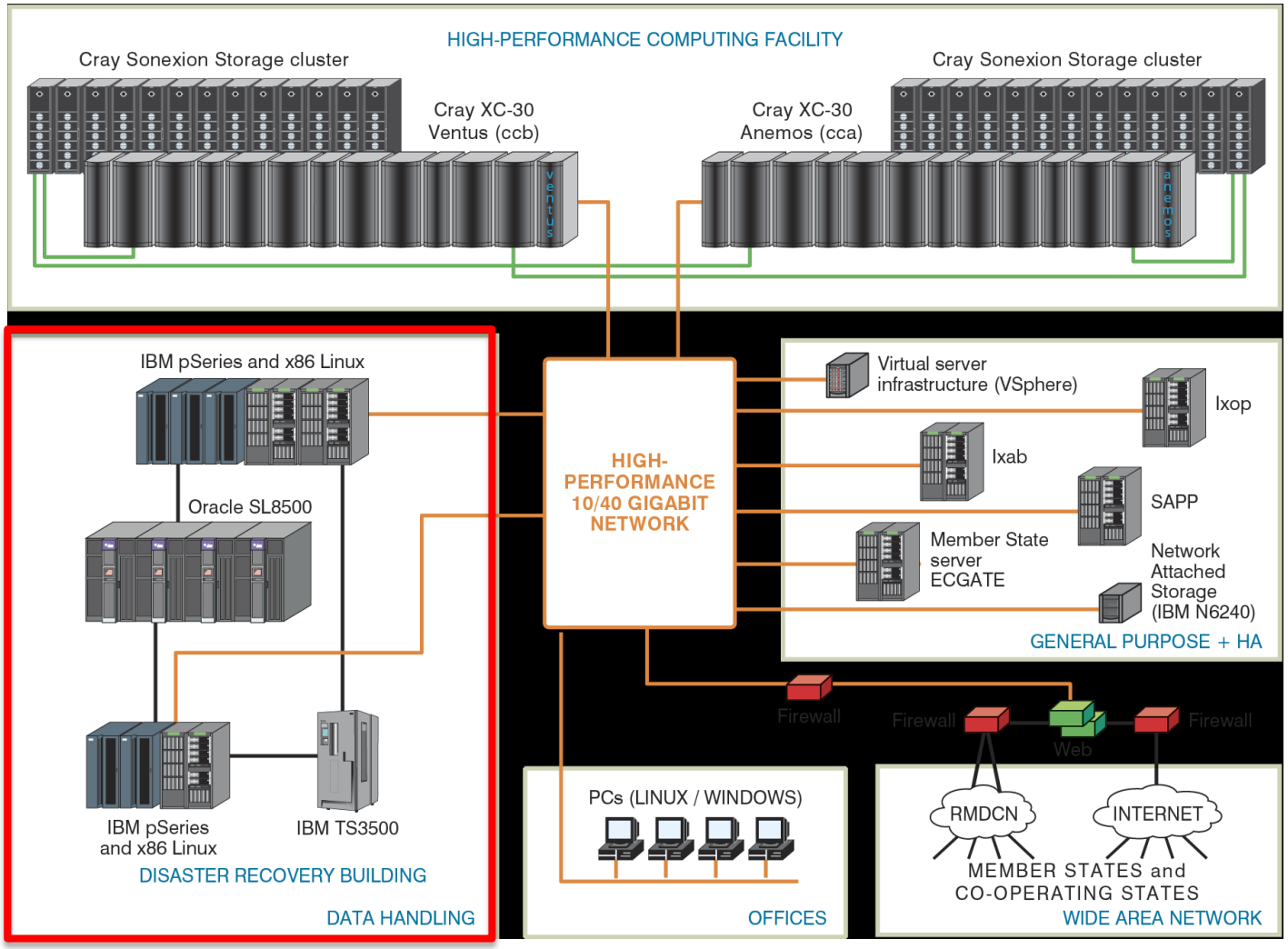


Lustre clusters:
About 14PB (combined)
Effective capacity: 10PB
Currently in use: 5PB

84000 cores.
Combined Power:
3600 teraflops (peak),
200 teraflops (sustained)

Data Handling System:
110 PB of data

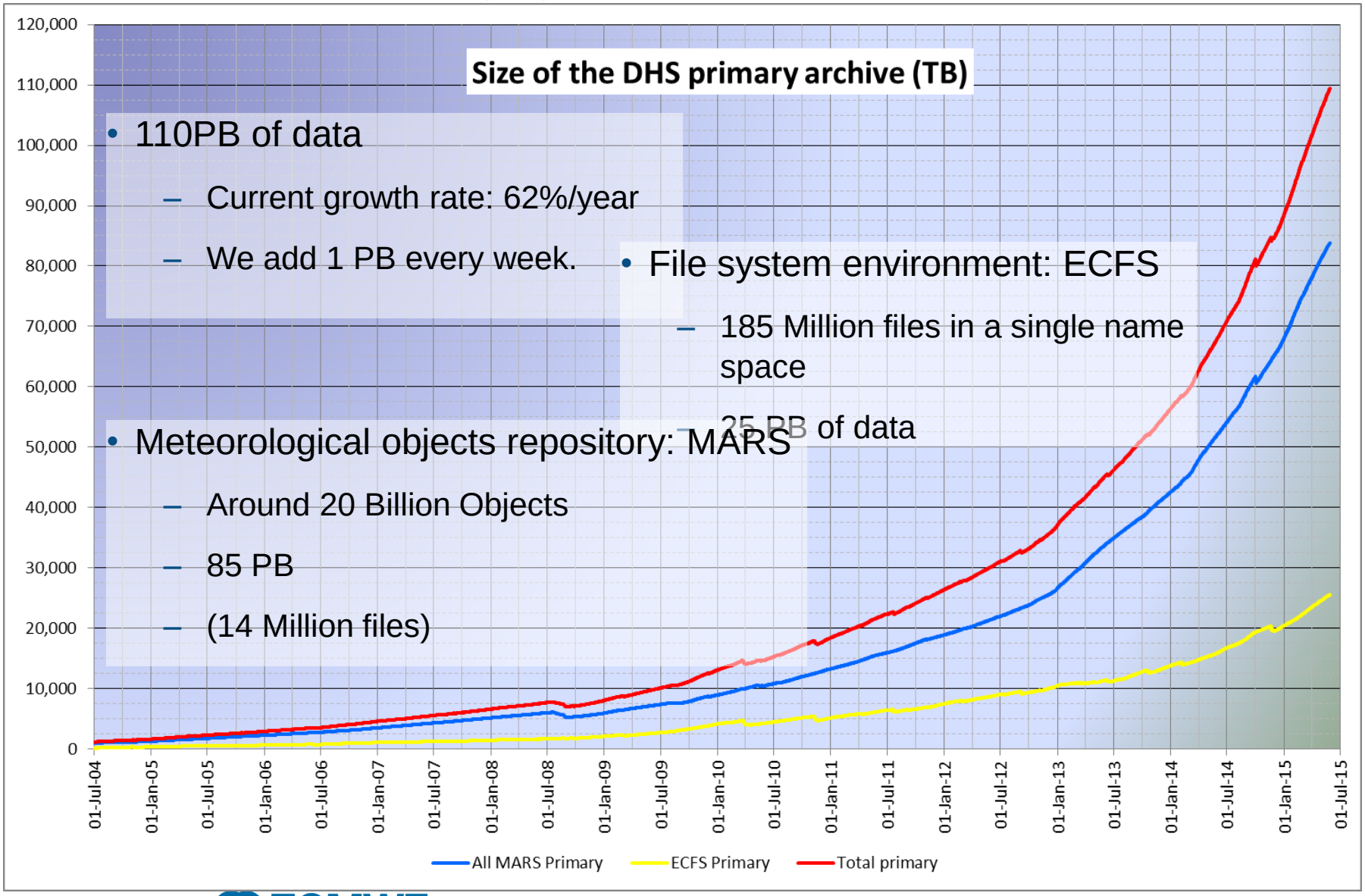
The Data Handling System



Our users

- Locally
  - Mainly HPC applications
  - Local scientists (-+200)
- Remotely
  - Member States Users
  - Meteorologists across the world
  - Thousands of registered users.
- Access type is quite varied. For example :
  - High resolution weather information (e.g. state of the atmosphere at a given time)
    - Done mostly locally
  - Access to higher level abstractions or lower resolution products
    - Both locally and remotely

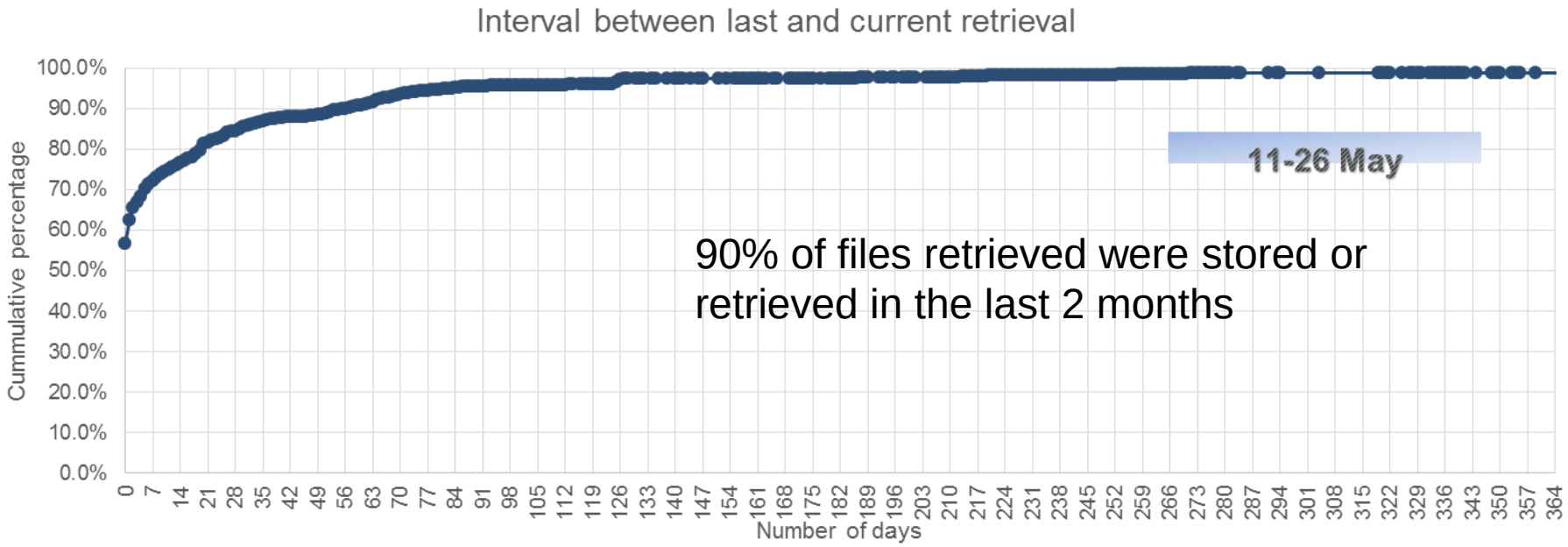
Our archiving system



- 110PB of data
  - Current growth rate: 62%/year
  - We add 1 PB every week.
- File system environment: ECFS
  - 185 Million files in a single name space
  - 25 PB of data
- Meteorological objects repository: MARS
  - Around 20 Billion Objects
  - 85 PB
  - (14 Million files)

Access Patterns: ECFS

Avg day store	Avg day retrieve	Peak hour store	Peak hour retrieve
40TB/day	20TB/day	2.6TB/H	2TB/H
500MB/s	250MB/s	750MB/s	600MB/s
150,000 files/day	60,000 files/day	15,000 files/hour	13,000 files/hour



Access Patterns: MARS

Avg day store	Avg day retrieve	Peak hour store	Peak hour retrieve
100TB/day	40TB/day	7.6TB/H	5TB/H
1200MB/s	500MB/s	2200MB/s	1500MB/s
150,000,000 objects/day	100,000,000 objects/day	10,000,000 objects/hour	7,500,000 objects/hour

- Some datasets are very popular. (less than 1%, .5PB)
- Operational datasets represent about 15PB,
  - Active for a few weeks
  - Then accessed from time to time
- Research experiments data: (70PB)
  - Typically dormant after a few days.
  - Some can become very popular after months or years

Behind the scene

- Applications:
  - Home grown:
    - ECFS (ECMWF File Storage)
    - MARS (Meteorological Archival and Retrieval System)
  - Backbone: HPSS
    - High Performance Storage System
    - Highly scalable Hierarchical Storage Management system
- Hardware used.
  - Servers:
    - Mostly small intel Linux servers.
  - Disk cache:
    - Midrange disk solutions : about 2.5PB of usable capacity
  - Tape storage:
    - 16,000 High-End tape media , providing 110PB of capacity.
    - 11,000 LTO media. (Safety copy of our most precious 20% of data)

We are still using tapes?

Hard to manage ?

Slow ?

Unreliable ?

Tapes usage at ECMWF: a few facts

- All data stored on tape in the days following its creation

- Only 10 to 20% of DHS user retrieval requests requires staging from tape to disk.
  - Still up to 15,000 tape mounts /day

- Only a small fraction of a tape contents is retrieved
  - (typically less than 1%)

- Access is very demanding:
  - small reads interspaced with positioning requests.

- Some tapes have been mounted > 10,000 times.

Why using tapes?

- Why not having 100PB of SSD or Disks?
- Tape is still significantly cheaper than those.
- Our budget does not grow exponentially.
- At constant budget, we could only sustain 30-50% annual growth rate, even with tape.
  - Our users want much more.
    - 60% this year, more in the future.

Direct access to all data
but
Reduce significantly the amount of data stored?
Loose potentially invaluable and irreproducible data?

50% growth rate (or more)
but
Delayed access for some retrieval requests?

But also…

- Reliability.
  - Data loss is extremely rare, typically self contained.
    - More than 80% of our data has no backup.
  - Less vulnerable to firmware/software induced corruption
- Performances are more than adequate for our needs.
  - Currently 200 drives, each able of 240MB/s
    - 48GB/s potential bandwidth.
    - In fact, we use much less than that
- Eye catching piece of equipment during site visits.

There are drawbacks…

Tape drives related Issues

- Latency
  - Less important in a predominantly batch environment.
  - Can be mitigated.
- Mechanical equipment
  - Tape drive and robotic failure do happen
- Requires supervision
  - One full time employee

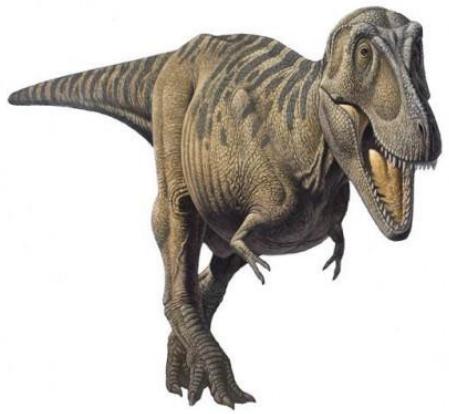
- Integration in “new” storage paradigms
  - Limited support in object storage solution.
- Tape capacity growth is slowing down
  - Used to be 40%/year, No more

Human related issues

- Tape mount storms
  - Tape drive thrashing
  - A badly defined requests can result in hundred requests.

- Users expectations need to be tailored
- Users need education.

- Finding the right skills to handle HSMs.
  - Hard to find manpower able to handle HSMs
    - What is an HSM?
  - Not seen as sexy.
  - Limited direct rewards.



How do we optimise our HSM?

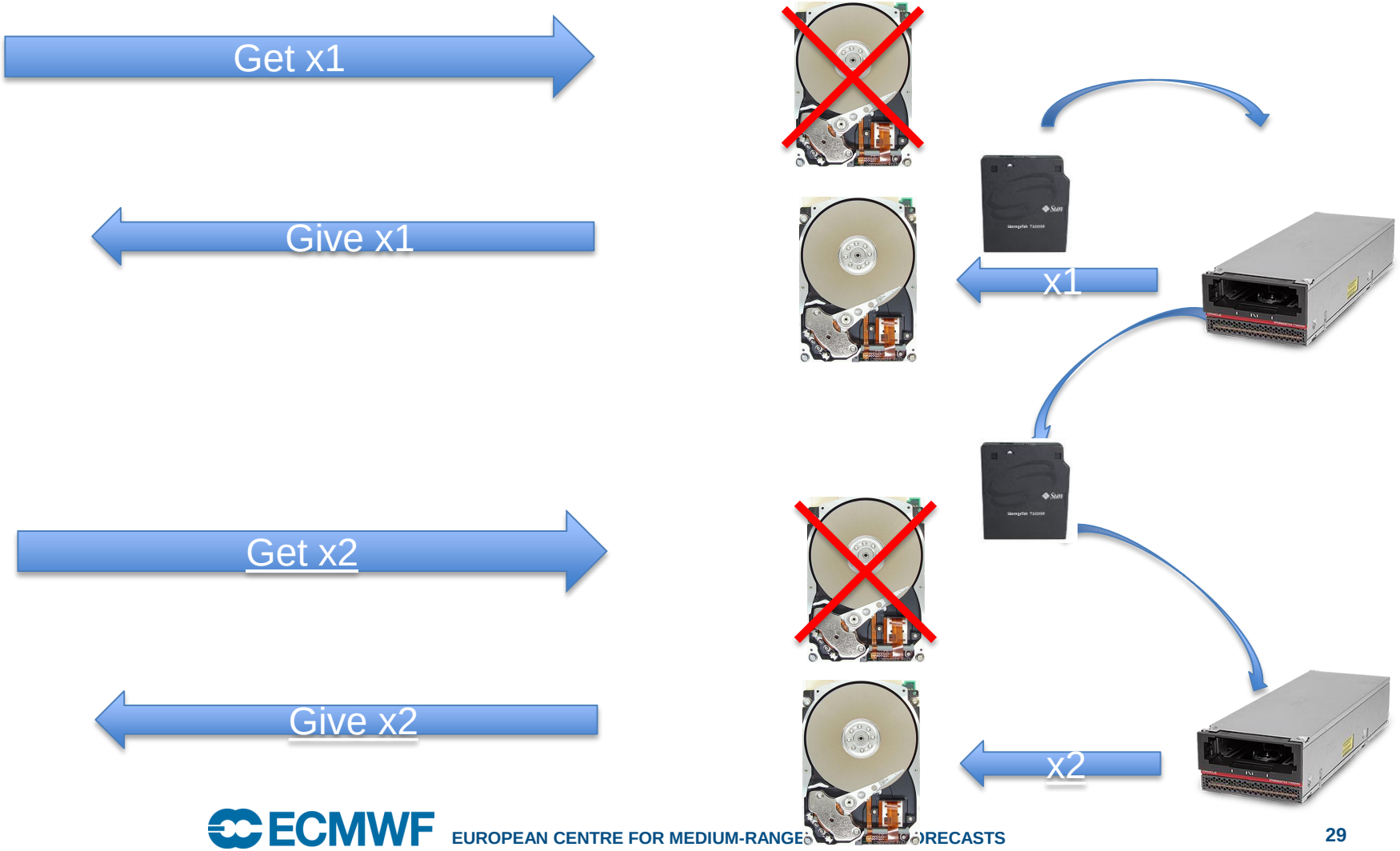
Good management of top storage cache!

Top Tier storage:
3% of the ECMWF archive
>85% of all retrieves

- ECFS:
  - Segregate files in different size bands.
  - In each band use Purge based on LRU.
  - Keep small files on disk forever/much longer than big ones.
- MARS:
  - Lock on disk very popular datasets (1/2 a PB).
  - Keep recently accessed objects for a few days on disk.
    - How long depends on dataset type.
  - Monitor for tapes being regularly used
    - Pre-cache and lock their contents.

Make each tape mount count.

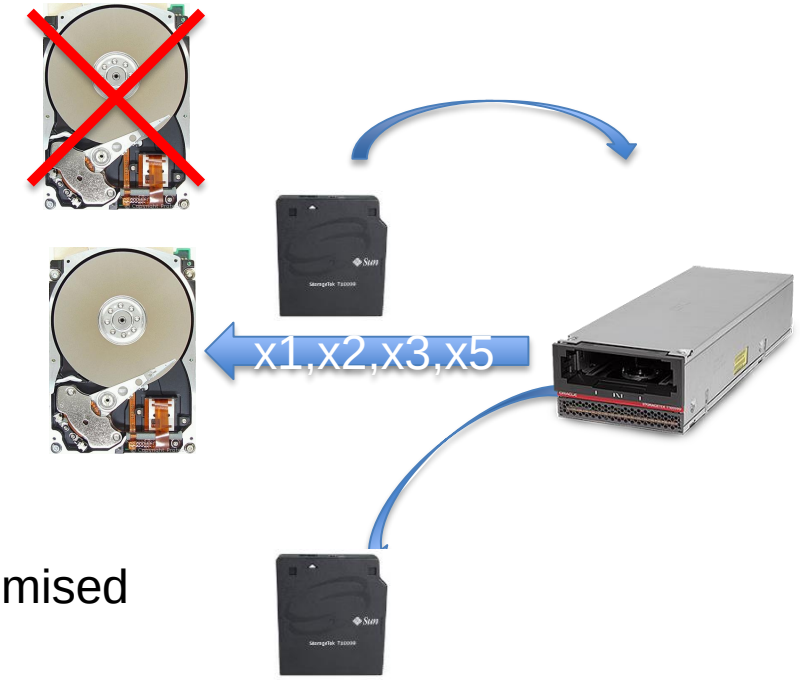
- Each Data request typically only retrieve a very small amount of data from tape.



Make each tape mount count.

- Reduce latency by asking multiple objects in one request

Get x1,x5,x2,x3



Give x1,x2,x3,x5

Requires user cooperation!

- Retrieval from tape can be optimised
  - Minimise positioning
  - New Tape Ordered Recalls features of several vendors should help become more efficient.

Application implementation dependant!

Make each tape mount count.

- Combining requests from multiple users.

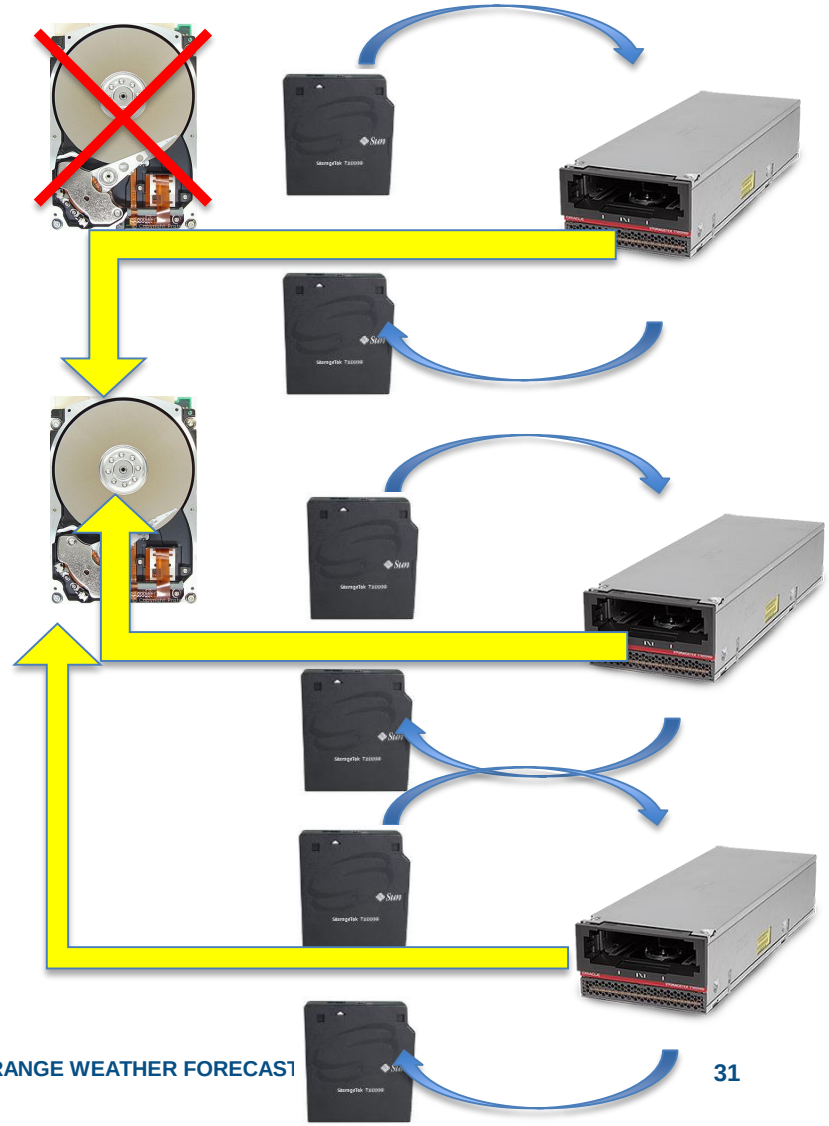
Get x1,x5,y3,x3,y7

Get x4,x5,z3

Give x1,x3,x5,y3,y7

Give x4,x5,z3

Optimisation done at application level

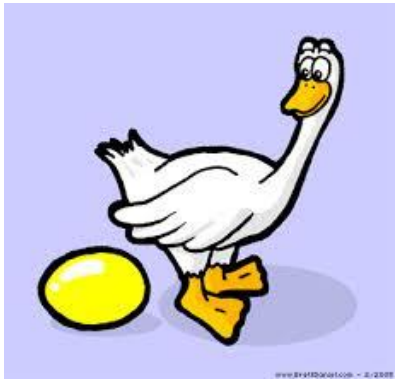


Use Enterprise-class drives.

- A lot of tape drives horror stories
  - Often linked to misuse of commodity equipment.
- Enterprise-class tape drives
  - Much better at positioning: more suitable to random access
  - Better at adapting to “low” IO bandwidth
  - Better at detecting and correcting errors
- Price differential is small.

When will HSM fade away at ECMWF?

- When we find an alternative solution …
  - Is cheaper than an tape-based HSM
  - Provide an aggregate bandwidth of tens of GiB/s
  - Allow access to first byte in <1 second in 90% of retrieves
  - Can be integrated in an environment presenting
    - Hundreds of millions of files in a single namespace.
    - hundreds of billions objects in a single repository
  - Can be protected against catastrophic failure
    - Access to data can be restored in a matter of hours.
  - Can be maintained/managed by a handful of people.
  - Provided by main stream vendor(s).



What are we exploring?

- Improve use of front end caches (Disks or SSD)
  - What do we gain by adding xx% of cache?
  - Better targeting of the caches to “likely candidates”..
- Use of new tape drives features
  - E.g. Tape Ordered Retrievals.
- We will look at erasure code based solutions.
  - Still need to find a reliable way to backup/restore these.

In the long run…

- Getting rid of file system paradigm?
  - File systems are popular with the users…
  - Implement usage of object-oriented storage solutions for MARS
  - Implement of object-accessible storage devices
  - Provide an “easy to use” database.
- Work is done in assessing ability to recreate data instead of conserving it.
  - Non trivial.
  - Impossible to recreate “the same” data.
- Revisit from time to time use of
  - dedup appliances and
  - Compression on disk.

In summary

- HSM allow us to keep more data for the same budget.

- Requires good management of higher storage tier
  - serve 80 to 90 % of requests from disk
  - Serve small files from disks

- Requires optimisation of tape usage.

- Very suitable for an environment where
  - A lot of batch usage
  - Users understand how to optimise their retrieves
  - Occasional wait for data is acceptable.

Any questions?

francis.dequenne@ecmwf.int

ECMWF