ECMWF’s Future challenges in Handling and Manipulating Model and Observational Data

Questions in “Big Data”

Tiago Quintino

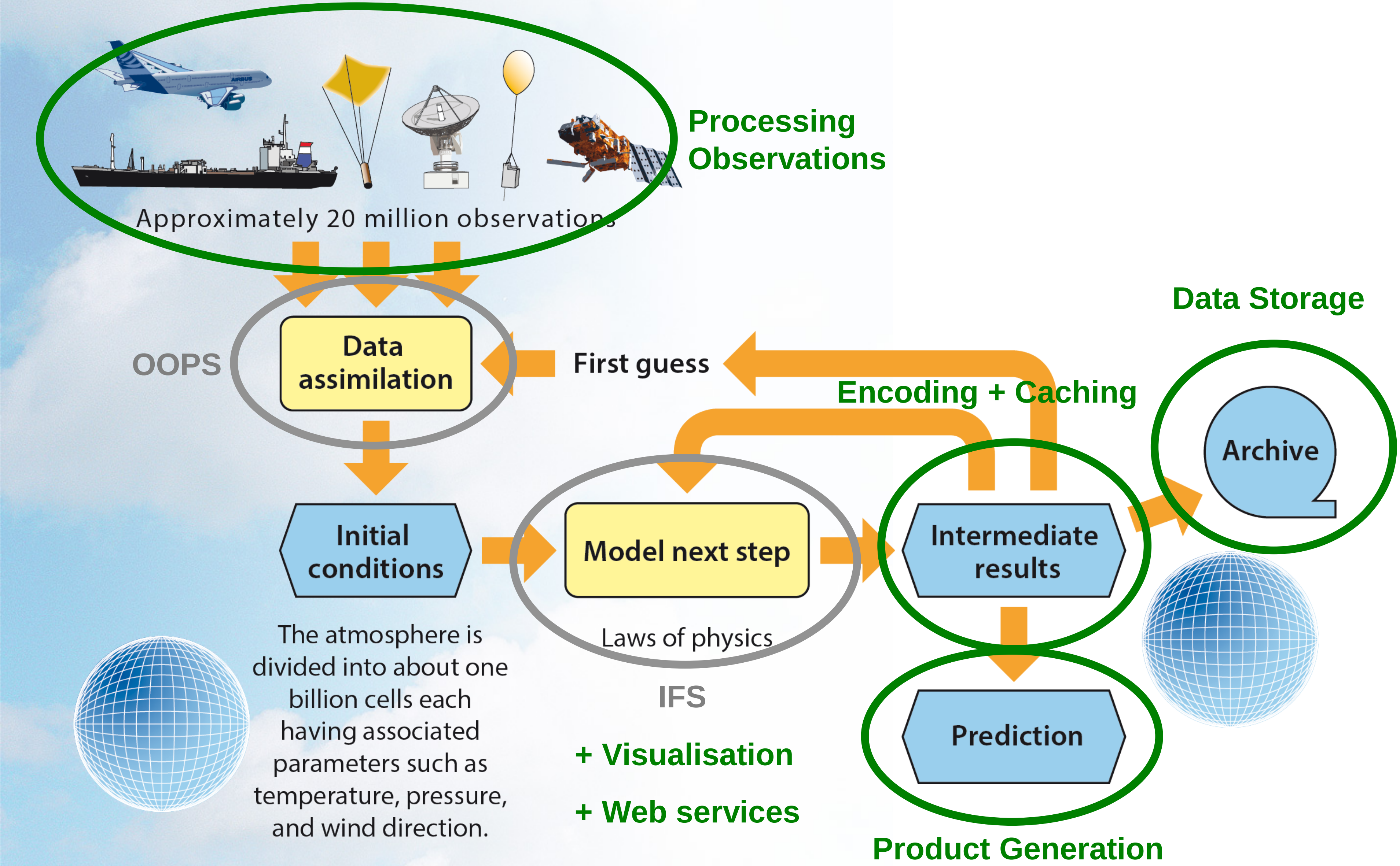
Data Handling

B. Raoult, M. Fuentes, S. Siemen

ECMWF

The Data Chain

A basic description of our models



Processing Observations

Approximately 20 million observations

OOPS

Data assimilation

First guess

Data Storage

Encoding + Caching

Archive

Initial conditions

Model next step

Laws of physics

Intermediate results

Prediction

The atmosphere is divided into about one billion cells each having associated parameters such as temperature, pressure, and wind direction.

IFS

+ Visualisation

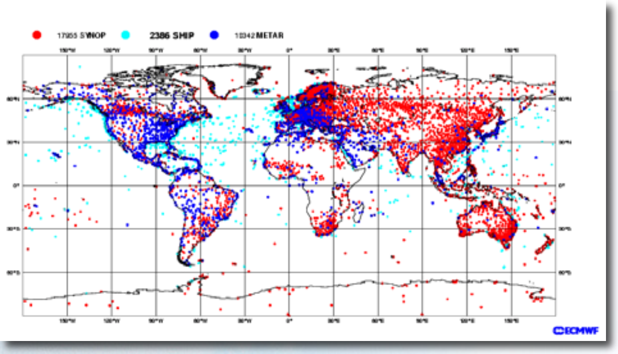
+ Web services

Product Generation

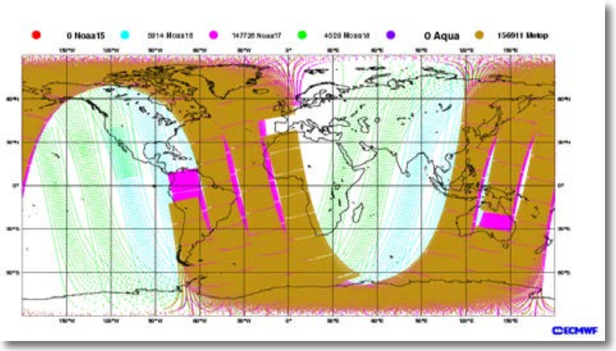
Major assimilated datasets

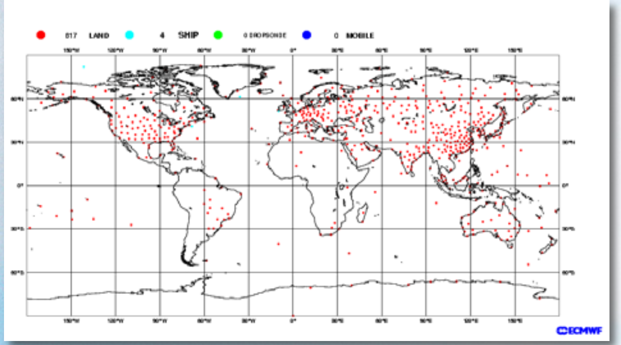
Receive 300 million observations from 130 sources daily.

Surface stations

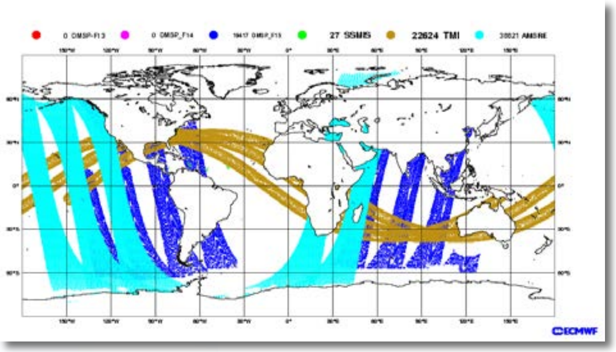


Polar, infrared



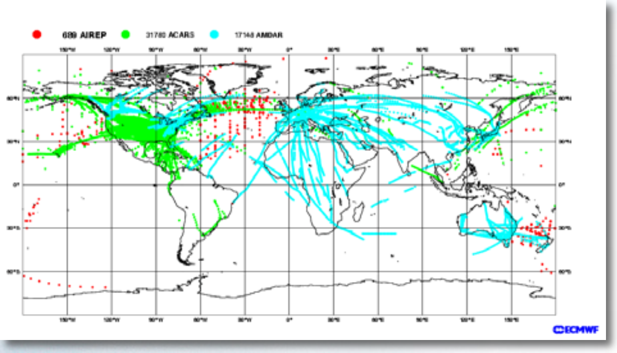


Radiosonde balloons

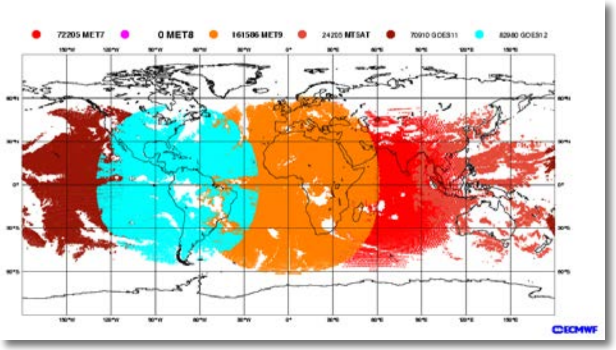


Polar, microwave

Aircraft



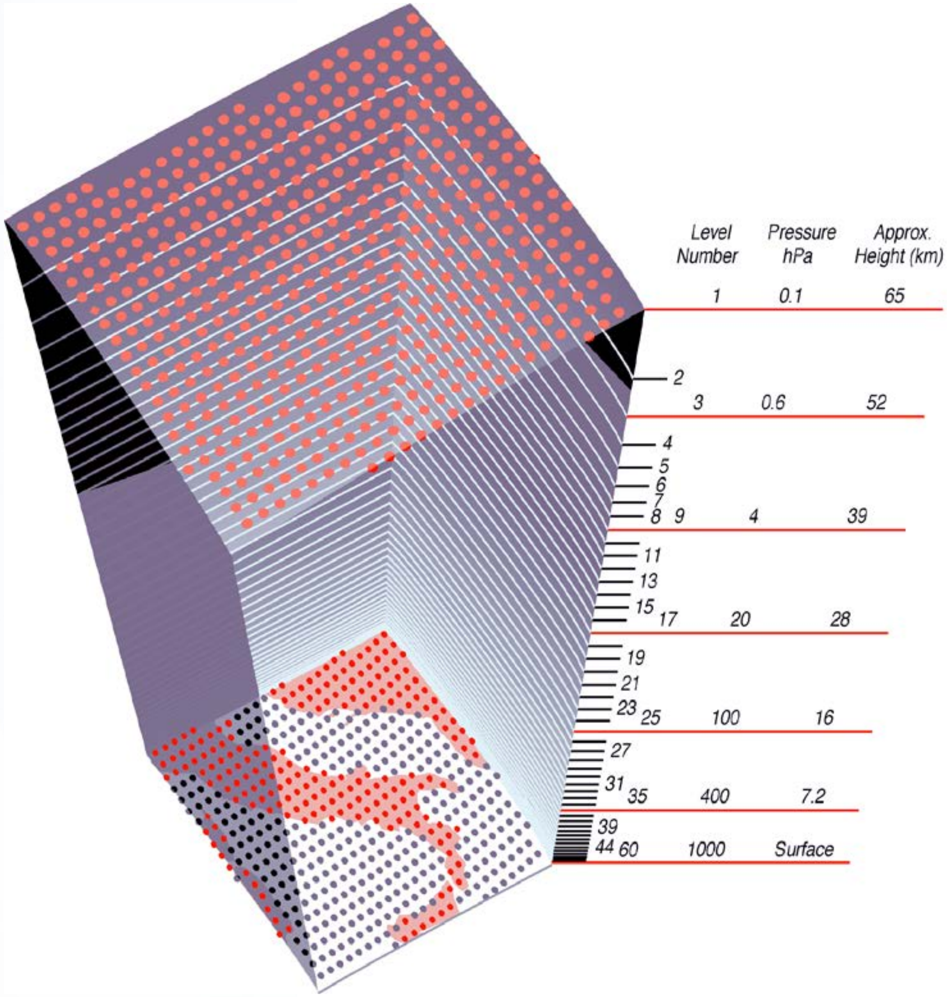
Geostationary, IR



Meteorological Fields

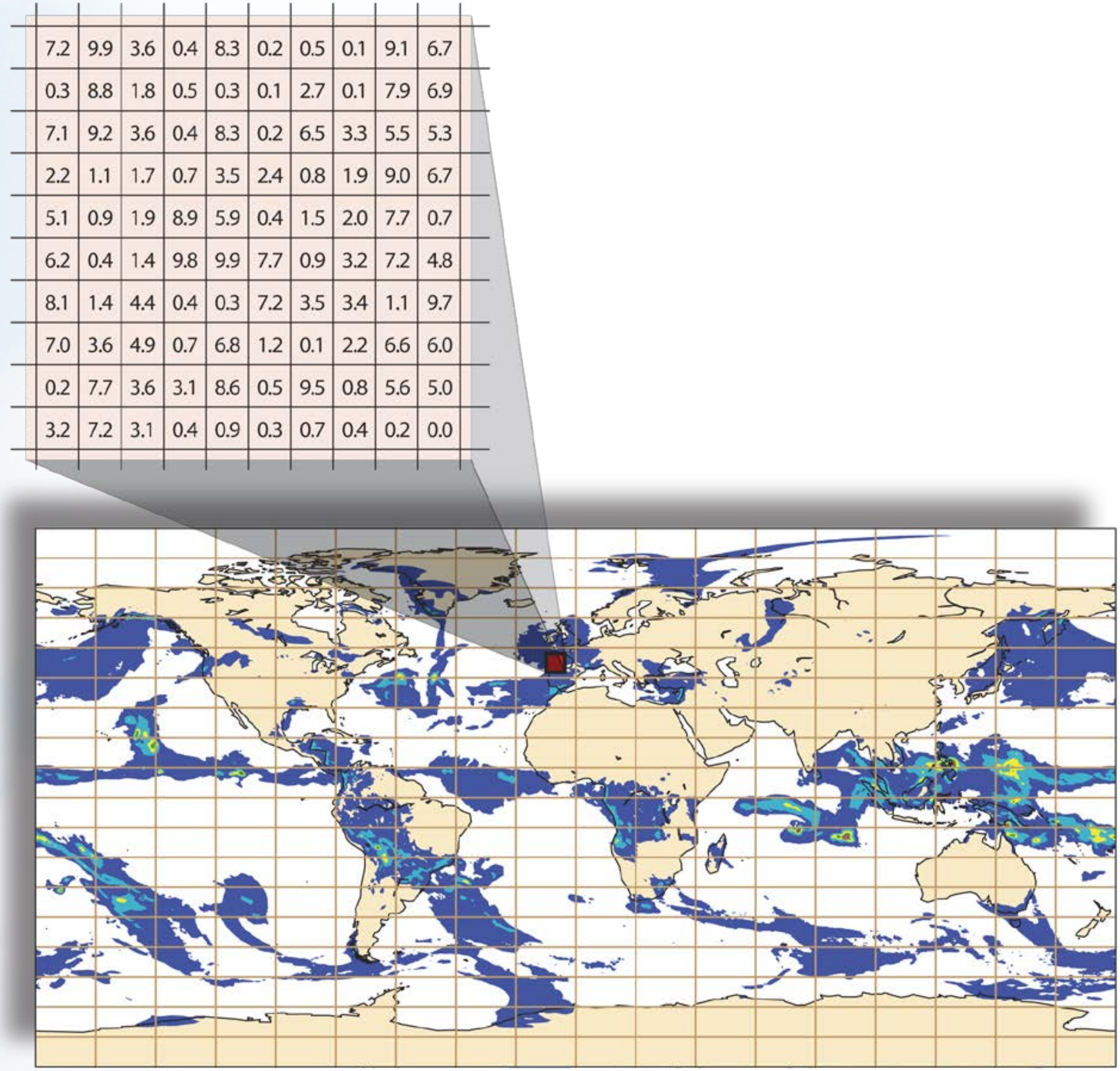
Operational models produce:

- 13 millions fields daily
- Totalling 8 TB/day



ECMWF products

- 77 million products disseminated ever day, totalling 6 TB.
- Interpolate output fields into **user required** grids
- Product generation is also subject to a **dissemination schedule** (time critical)
- Products also served via web **visualisation services**



7.2	9.9	3.6	0.4	8.3	0.2	0.5	0.1	9.1	6.7
0.3	8.8	1.8	0.5	0.3	0.1	2.7	0.1	7.9	6.9
7.1	9.2	3.6	0.4	8.3	0.2	6.5	3.3	5.5	5.3
2.2	1.1	1.7	0.7	3.5	2.4	0.8	1.9	9.0	6.7
5.1	0.9	1.9	8.9	5.9	0.4	1.5	2.0	7.7	0.7
6.2	0.4	1.4	9.8	9.9	7.7	0.9	3.2	7.2	4.8
8.1	1.4	4.4	0.4	0.3	7.2	3.5	3.4	1.1	9.7
7.0	3.6	4.9	0.7	6.8	1.2	0.1	2.2	6.6	6.0
0.2	7.7	3.6	3.1	8.6	0.5	9.5	0.8	5.6	5.0
3.2	7.2	3.1	0.4	0.9	0.3	0.7	0.4	0.2	0.0

Questions in "Big Data"

"There are no right answers to wrong questions"
- Ursula Le Guin

What is Big Data?

“Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization.”

“Big Data”, Wikipedia, retrieved 2014

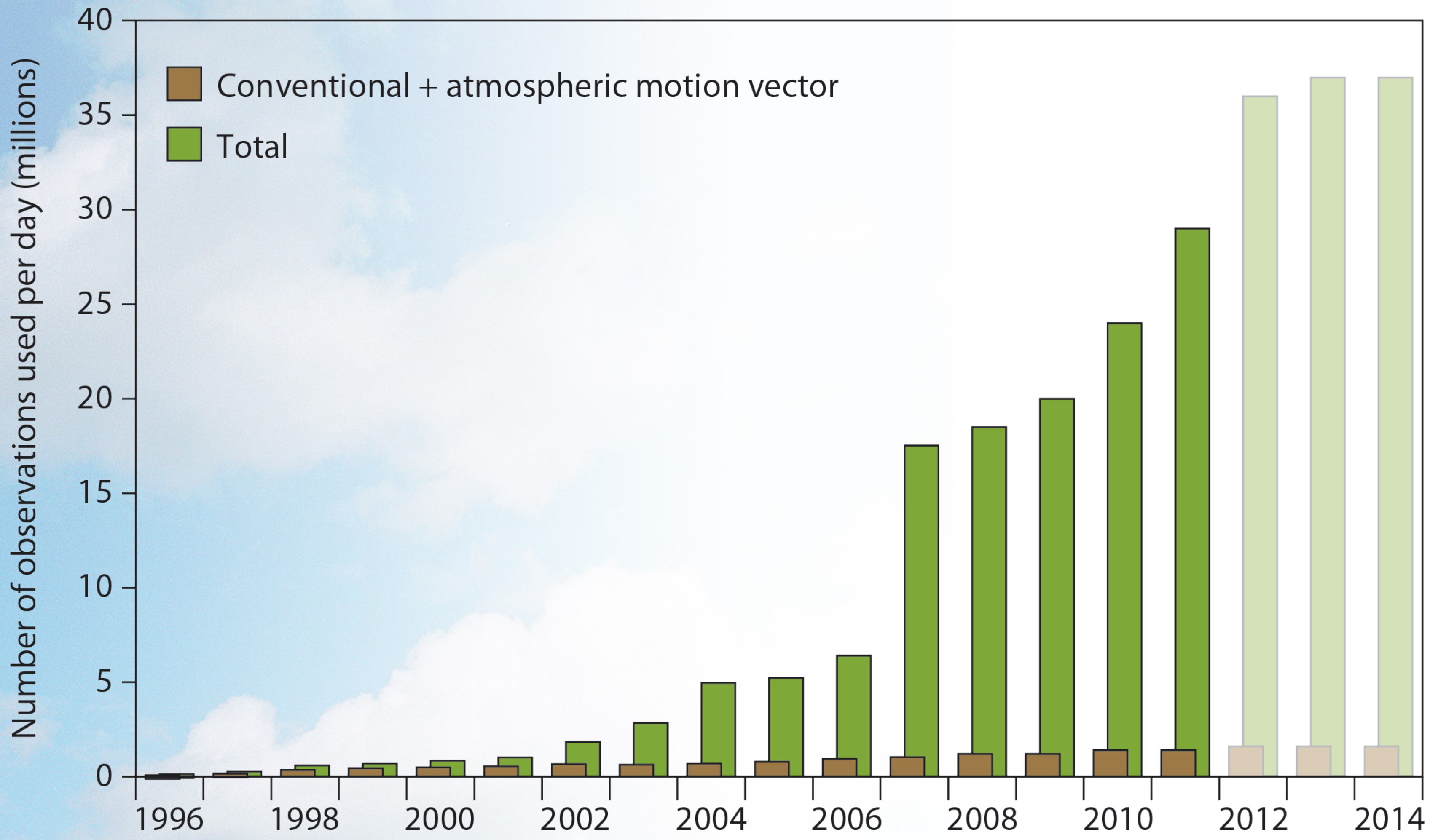
“Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”

“3D Data Management: Controlling Data Volume, Velocity and Variety”, D. Laney, Gartner, 2001

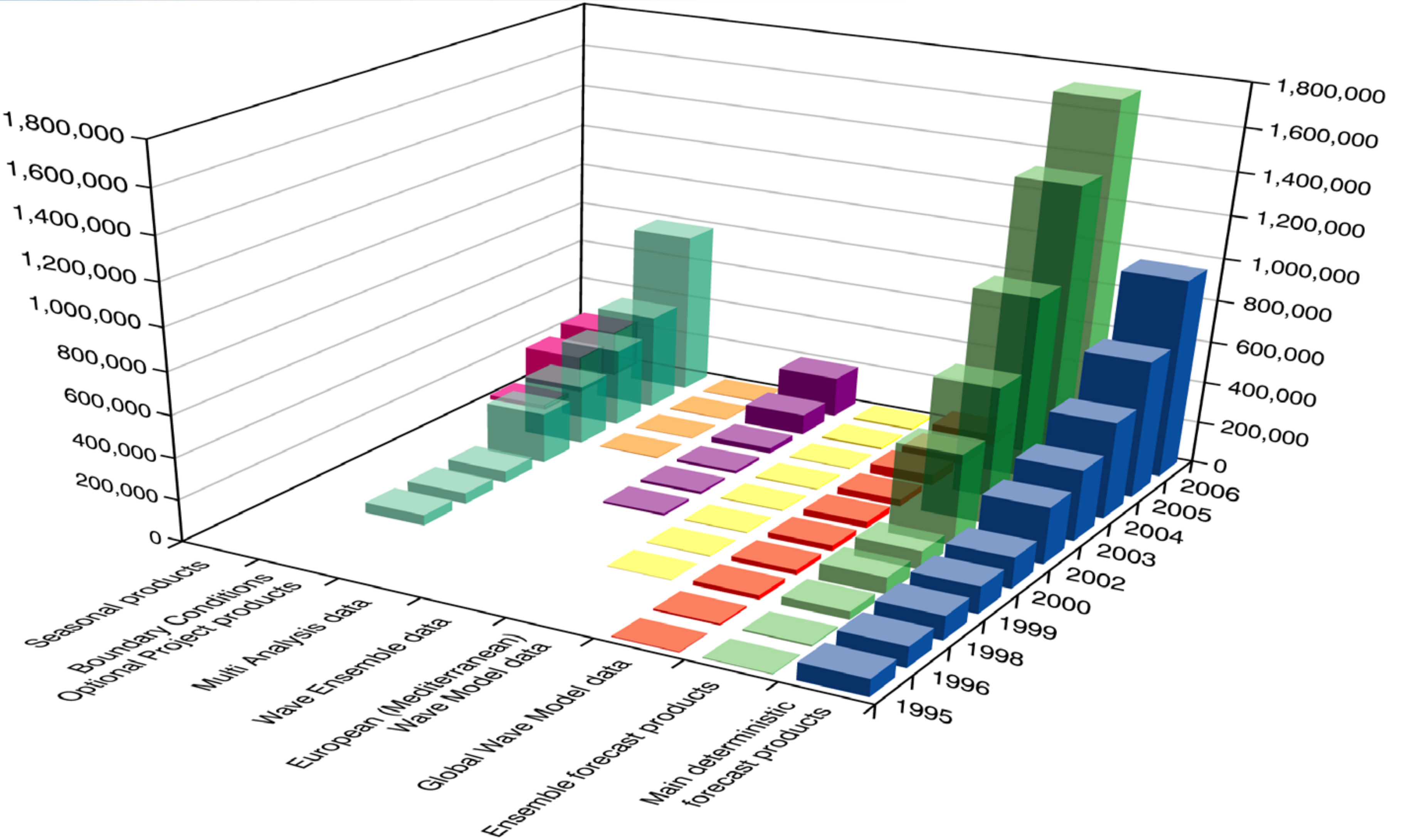
The 3 V’s of Big Data

V is for Volume

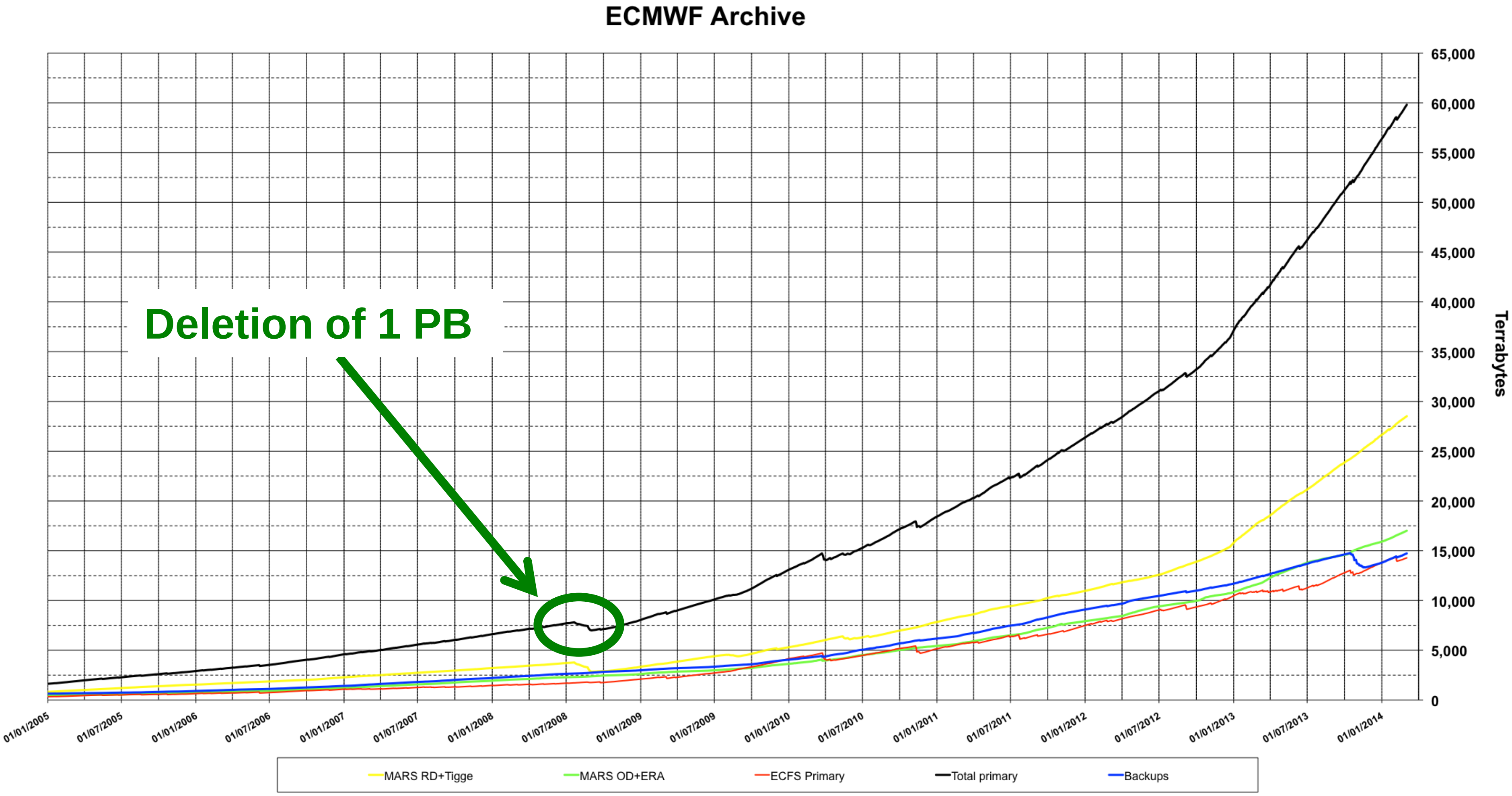
Increase of satellite data usage



V is for Volume



V is for Volume



V is for Velocity

- ECMWF’s archive grows exponentially:

Initial volume *Time*

$$V = V_0(1+r)^t$$

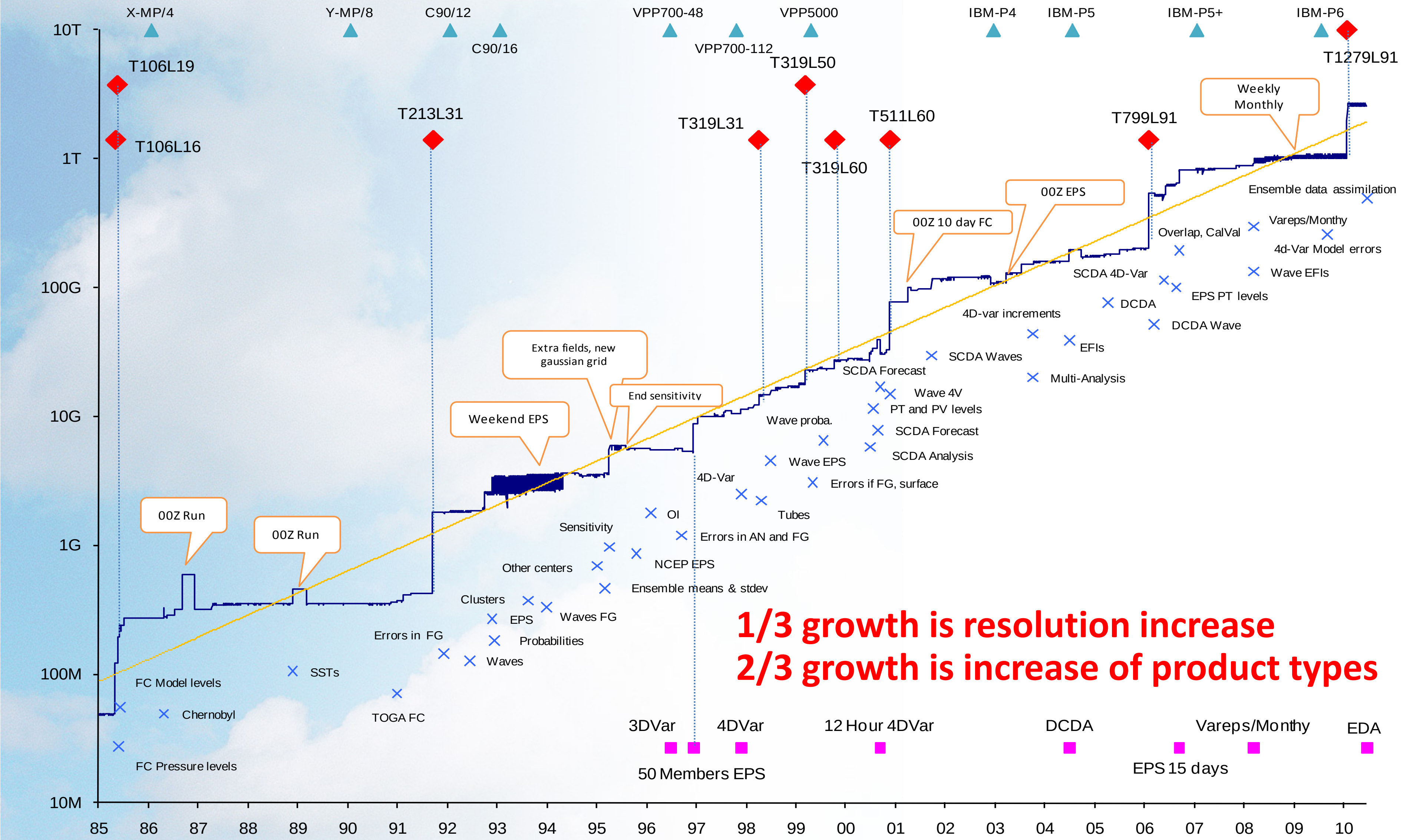
Volume of the archive *Rate of growth*

  - r is around 0.5, which is a 50% increase per year
  - The rate of added data also grows exponentially at the same rate!

$$\frac{\partial V_0(1+r)^t}{\partial t} = V_0 log(1+r)(1+r)^t = A_0(1+r)^t$$

- In 1995, the size of the archive was increasing at a rate of 14 TB/year.
- In 2014, the size of the archive increases at a rate higher than 65 TB/day with peaks of **100 TB/day**

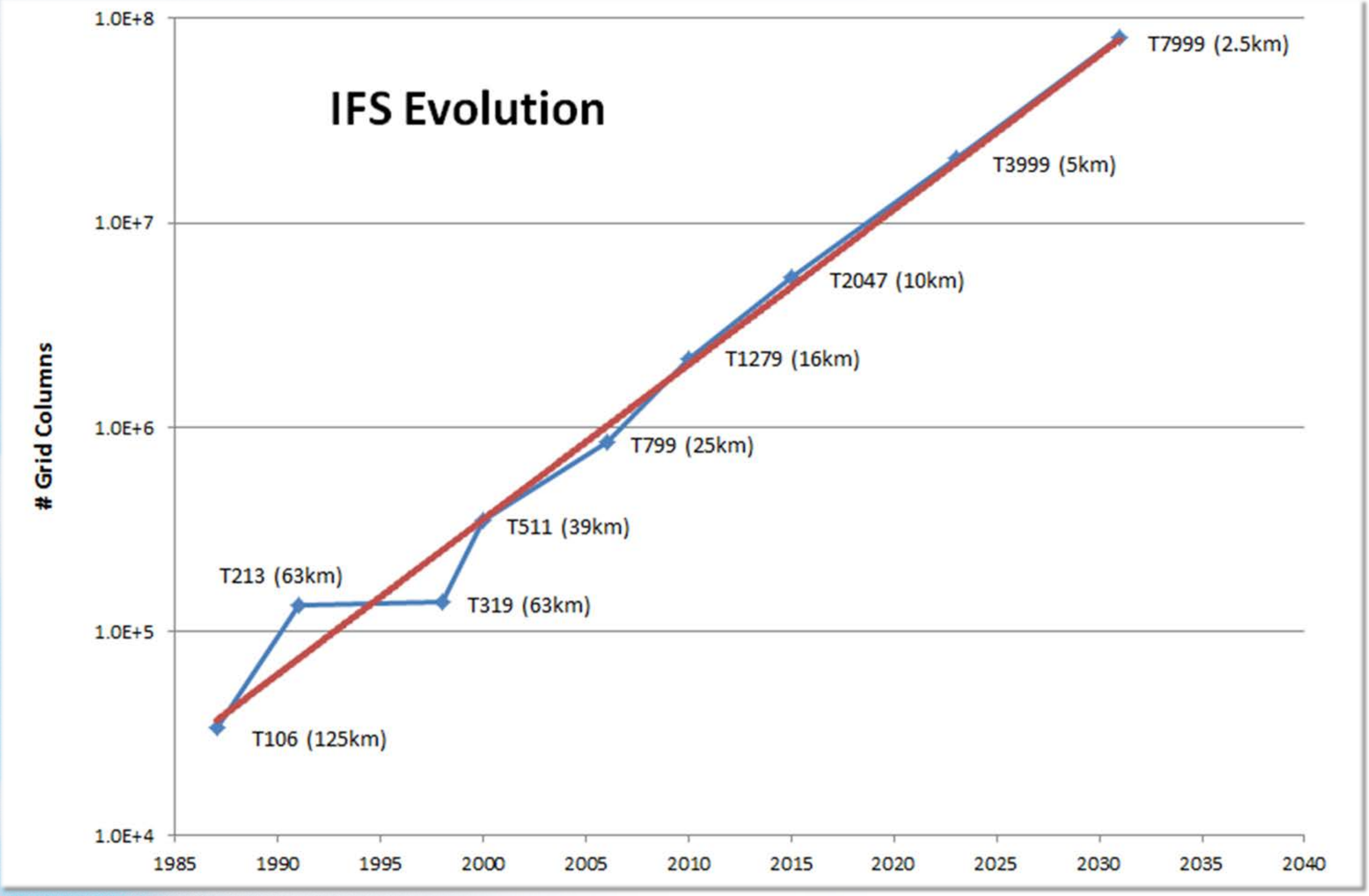
V is for Variety



1/3 growth is resolution increase
2/3 growth is increase of product types

Future Challenges

… more of the same?


IFS Evolution
# Grid Columns
1.0E+8
1.0E+7
1.0E+6
1.0E+5
1.0E+4
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
T7999 (2.5km)
T3999 (5km)
T2047 (10km)
T1279 (16km)
T799 (25km)
T511 (39km)
T213 (63km)
T319 (63km)
T106 (125km)

ECMWF

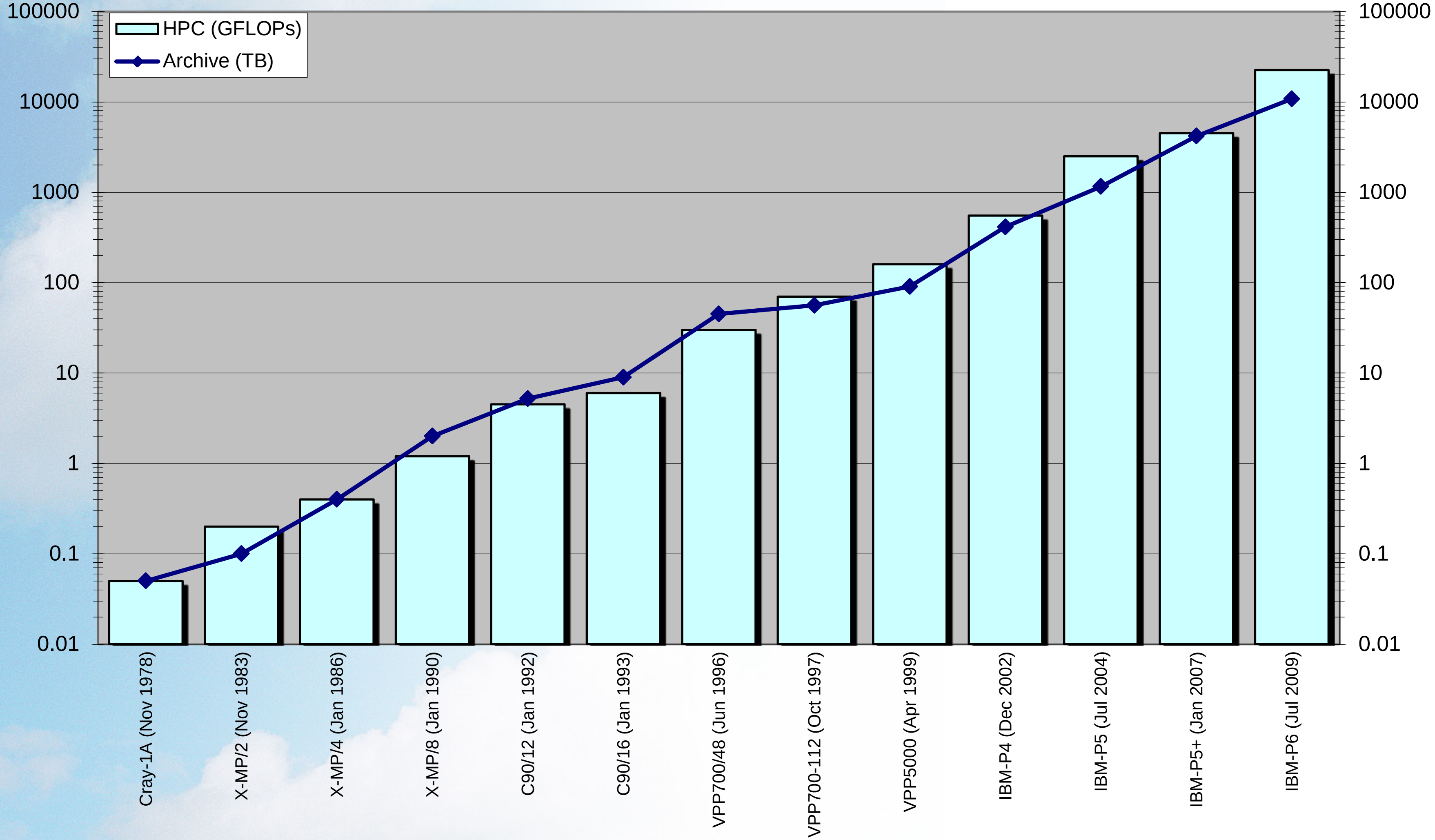
Impact of Resolution Upgrades

Resolution	Grid size	Grid Points	Field Size (in memory)
T319	62.5 km	204 k	1.6 MB
T511	39 km	524 k	4 MB
T799	25 km	1.2 M	9.6 MB
T1279	16 km	2.1 M	16.8 MB
T2047	*10 km*	*8.4 M*	***67.2 MB***
T3999	*5 km*	*20 M*	***160 MB***
T7999	*2.5 km*	*80 M*	***640 MB***

As memory per core diminishes (think GPU's) …

… this may have serious implications on the interpolation software!

Archive size vs. Supercomputer power



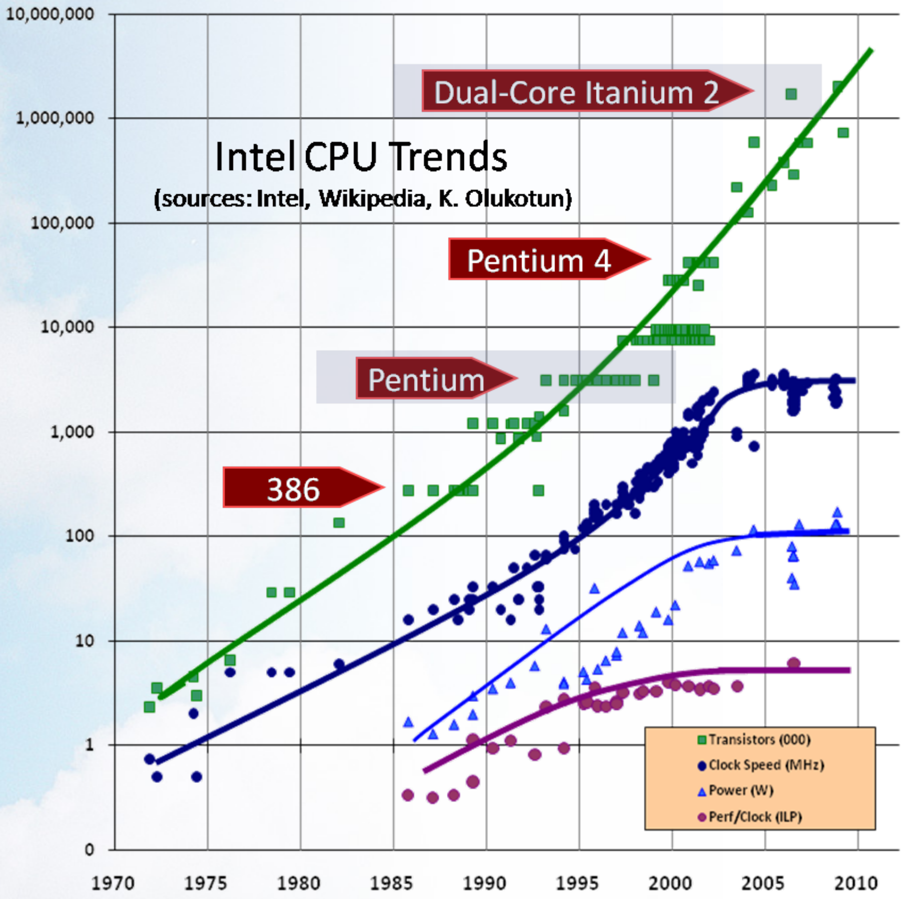
As has been impacting on the archive size…

Nothing of this is new
We have always been dealing with this issues…

What changed?

CPU Power Growth

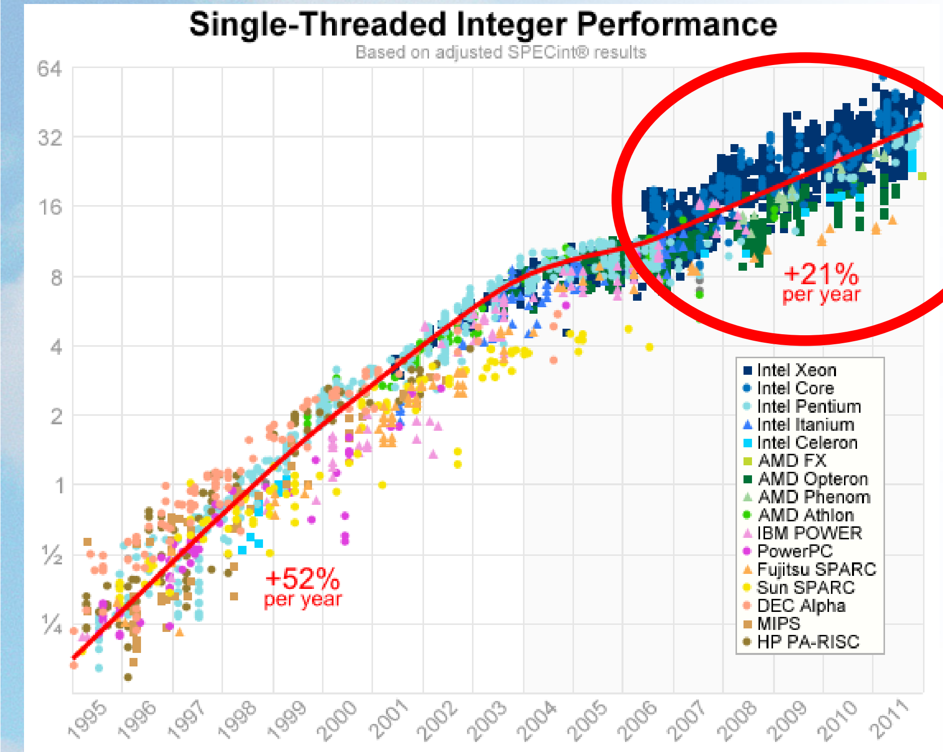
"The Free Lunch is Over". H. Sutter, Dr. Dobb's Journal, 30(3), March 2005



But what about "real" performance?

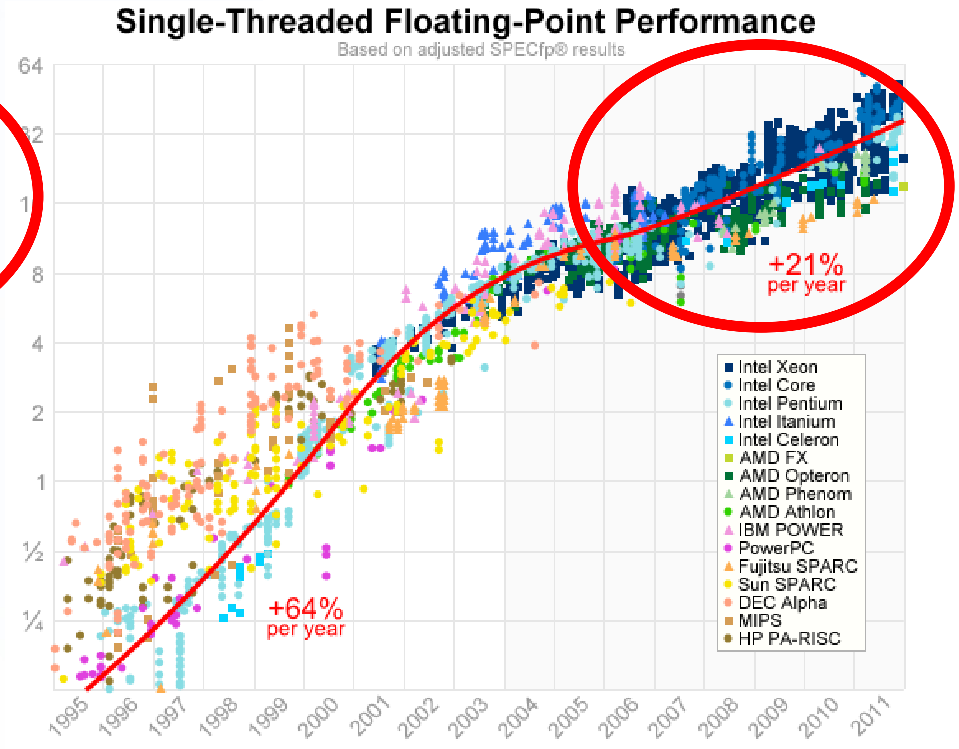
CPU Performance Growth (single-threaded)

"A Look Back at Single-Threaded CPU Performance", J. Pershing Feb 2012



Single-Threaded Integer Performance

Based on adjusted SPECint® results



Single-Threaded Floating-Point Performance

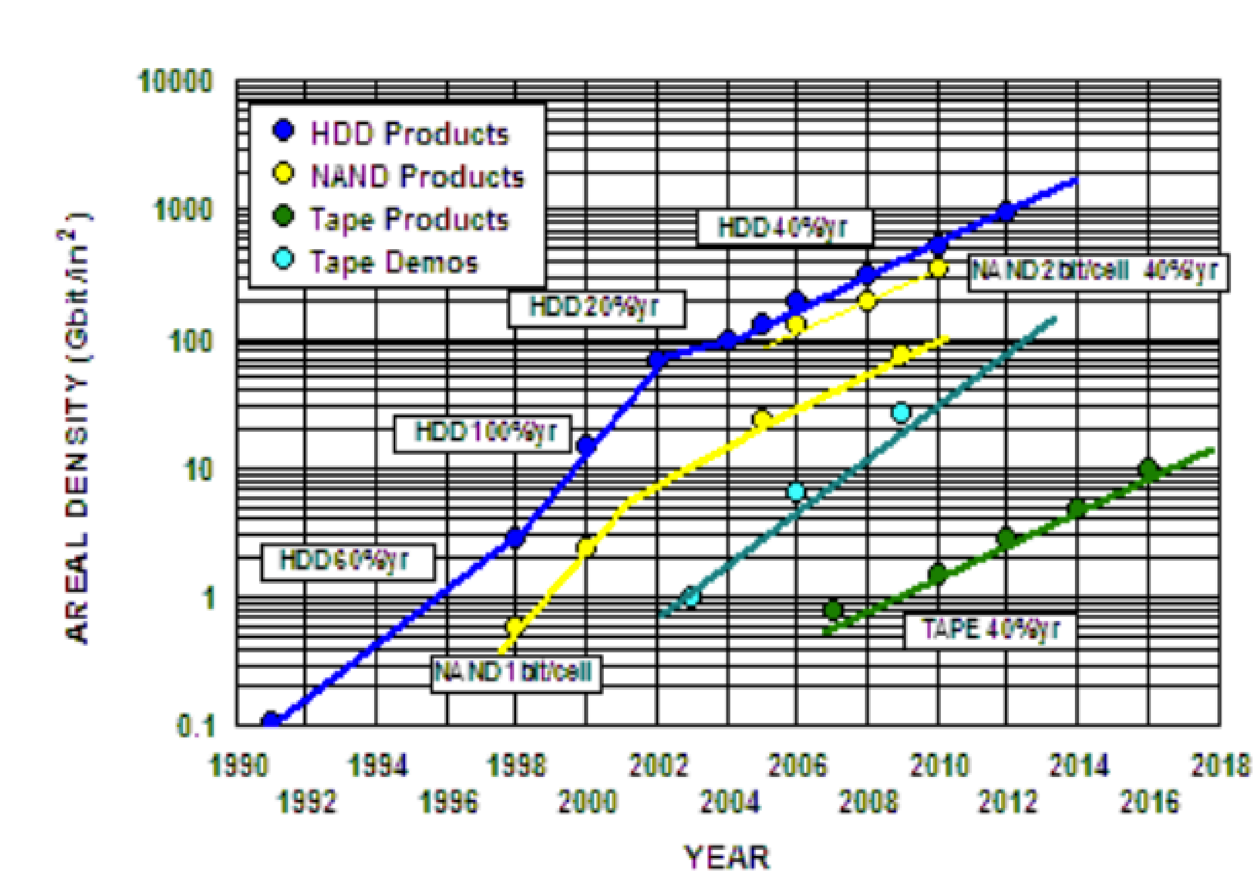
Based on adjusted SPECfp® results

More registers, vector units, branch prediction …

… but also harder to achieve!

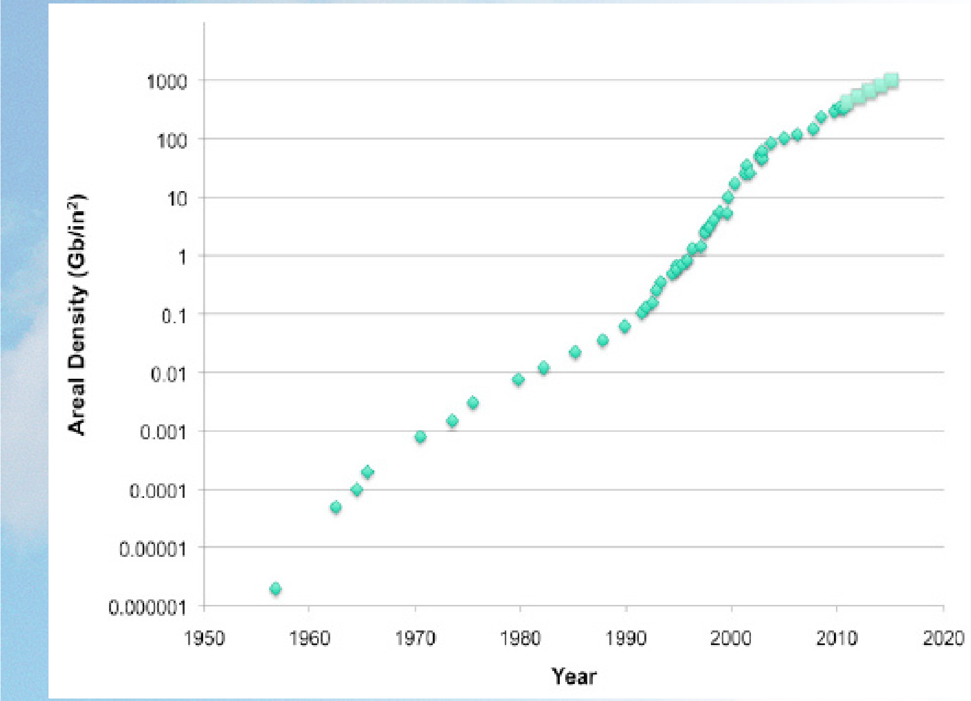
Storage Density Growth – Multiple Technologies

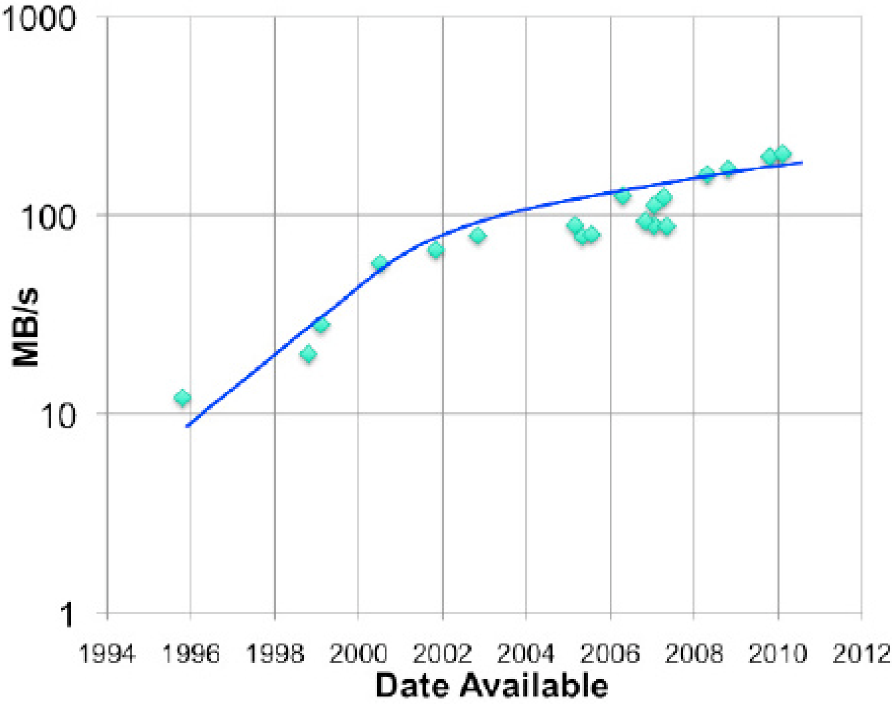
"Tape based magnetic recording: technology landscape comparisons with hard disk drive and flash roadmaps", R. Fontana et al, IBM Research Division, 2011



HDD Storage Growth

“GPFS Scans 10 Billion Files in 43 minutes”. R. Freitas, et al. IBM Research Division, 2011





Volume is linearly proportional to area density
Recently follows 25-40% CAGR…

… but transaction rate hasn’t kept up!

This means that we may have the capacity, but maybe not the bandwidth …

What does it imply?

- **“No Free Lunch”** ➔ Improve our software

- Explore new **Algorithms** that expose …
  - Concurrent computations (eg. map-reduce)
  - Data locality (eg. FEM discretisations)
  - Computational intensity (CPU usage/MB transferred)

- Software must cope with changes – **Flexibility**
  - Best use of new hardware (eg. use high-level DSL)
  - Unknown future for parallel platforms
  - Be able to adapt to changes in system architecture

Can we do it?

We have already started…

+ OOPS project for Assimilation

+ IFS Co-Array Fortran

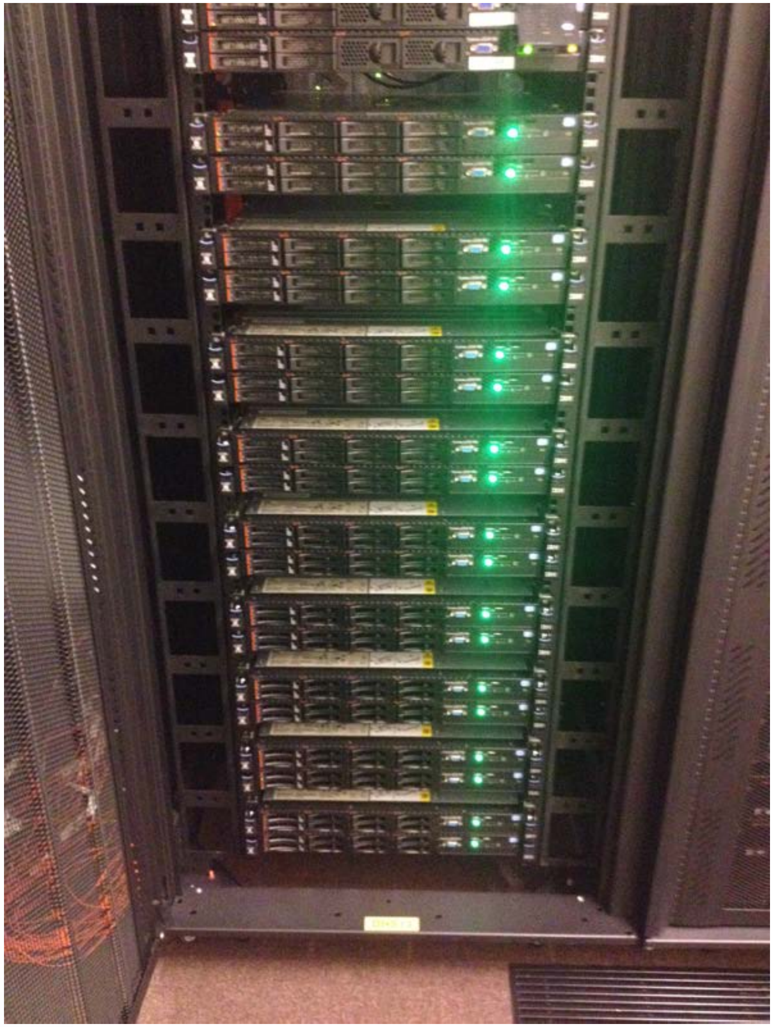
+ PantaRhei project

ECMWF's Meteorological Archival and Retrieval System

- A managed archive, **not a file system**
  - Users not aware of the location of the data
  - Retrievals expressed in meteorological terms
- Data is kept **forever**:
  - Dataset becomes more useful once enough data has been accumulated
  - Deleting old data in an exponentially growing archive is meaningless
- Consists of 3 layers:
  - FDB - cache at the HPC level (~80% hit ratio)
  - DHS - HDD cache (~80% hit ratio)
  - HPSS Tape system

ECMWF's Meteorological Archival and Retrieval System

- **Fully distributed** (migrated 2012)
  - 15 servers for metadata and data movers
  - 40 PB primary archive
  - 1 PB of disk cache (2.5%)
  - 110 billion fields in 8.5 million files
  - 200 million objects/65 TB added daily
  - 7000 registered users
  - 650 daily active users
  - 100 TB retrieved per day, in 1.5 million requests



# Users and # Requests **not** directly under our control…

➔ Scale with # Users / Requests !

A meteorological language

-
  ```
  retrieve,
  date        =      20110101/to/20110131,
  parameter   =      temperature/geopotential,
  type        =      forecast,
  step        =      12/to/240/by/12,
  levtype     =      pressure levels,
  levels      =      1000/850/500/200,
  grid        =      2/2,
  area        =      -10/20/10/0
  ```

- This request represents 31*2*20*4 = 4960 fields

Indirection is key to Scalability

IFS I/O Layer

As IFS improves its scalability …

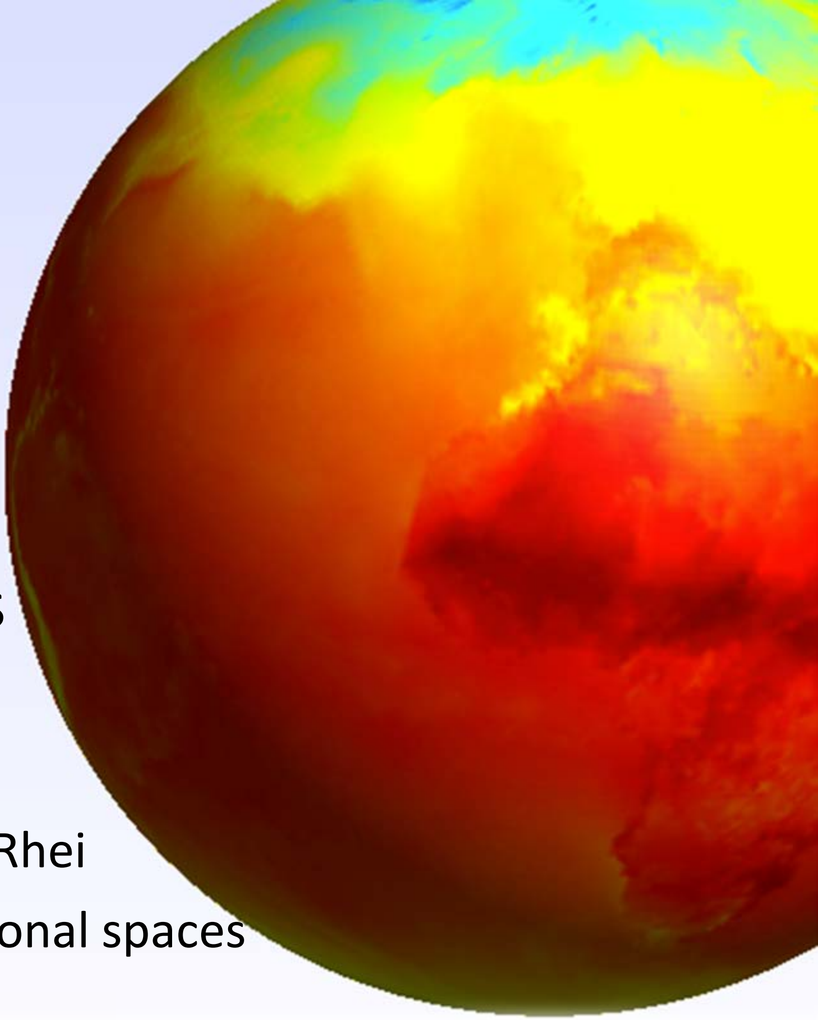
- GRIB encoding is likely to become a bottleneck
  - GRIB encoding requires full field (involves data gather)
  - Currently done within **IFS**

➔ **Introduce an I/O layer (indirection)**

- Achieve **adaptability** to changing paradigms:
  - Do data gather on our side?
  - Implement IO Server?
  - Encode GRIB in parallel? Defer encoding?
  - Encode in a parallel format (NetCDF4? Other?)

➔ Very important to optimize the **whole** data chain 🡐

Other Data Chain Components



Currently under development …

- Observations
  - COPE project: real time processing
- IFS I/O (cached storage)
  - FDB5: transactional & integrated with MARS
- Interpolation and Product Generation
  - New interpolation package (MIR)
  - ATLAS Framework co-developed with PantaRhei
  - Looking into FEM data-structures and functional spaces

Needing future attention …

- Visualisation
- Encoding fields (GRIB, NetCDF)

Summary

- ECMWF Data Chain faces **the Big Data 3V's** scalability challenges…

- Need to develop **concurrent** approaches to **all** data chain components:
  - Observation Processing
  - Data Encoding
  - Data Storage
  - Interpolation and Product Generation
  - Visualisation

- I/O transaction rates are not keeping up with growth
  - Avoid I/O by pipelining between data-chain components?
  - Move processing closer to the data?
  - Meteorology "Cloud Services"?

Shameless Advertising

We are **hiring** !

Visit www.ecmwf.int > Employment

- **Scalability** Program
- Work in the Data Handling Team

- Looking for experts in:
  - High Performance Computing
  - GPU's, Accelerators
  - Algorithms

Come and help us solve these challenges …

Questions?

* No dwarfs were used in the production of this presentation

** OK, except maybe one called MapReduce…

“The Landscape of Parallel Computing Research: A View from Berkeley”, Asanovic et al, December 2006 (aka 13 Berlekey Dwarfs)