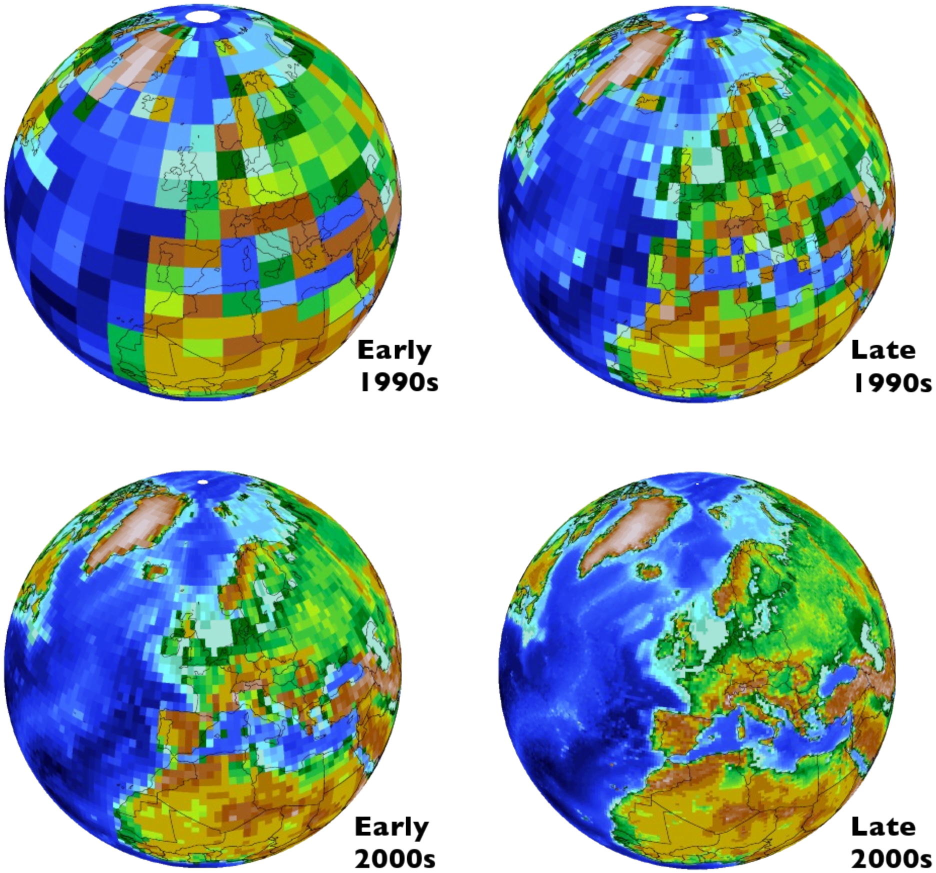
Cyberinfrastructure Challenges (from a climate science repository perspective)

Bryan Lawrence

CEDA

Rutherford Appleton Laboratory

Simulation Data Deluge



Globes courtesy of Gary Strand (NCAR)

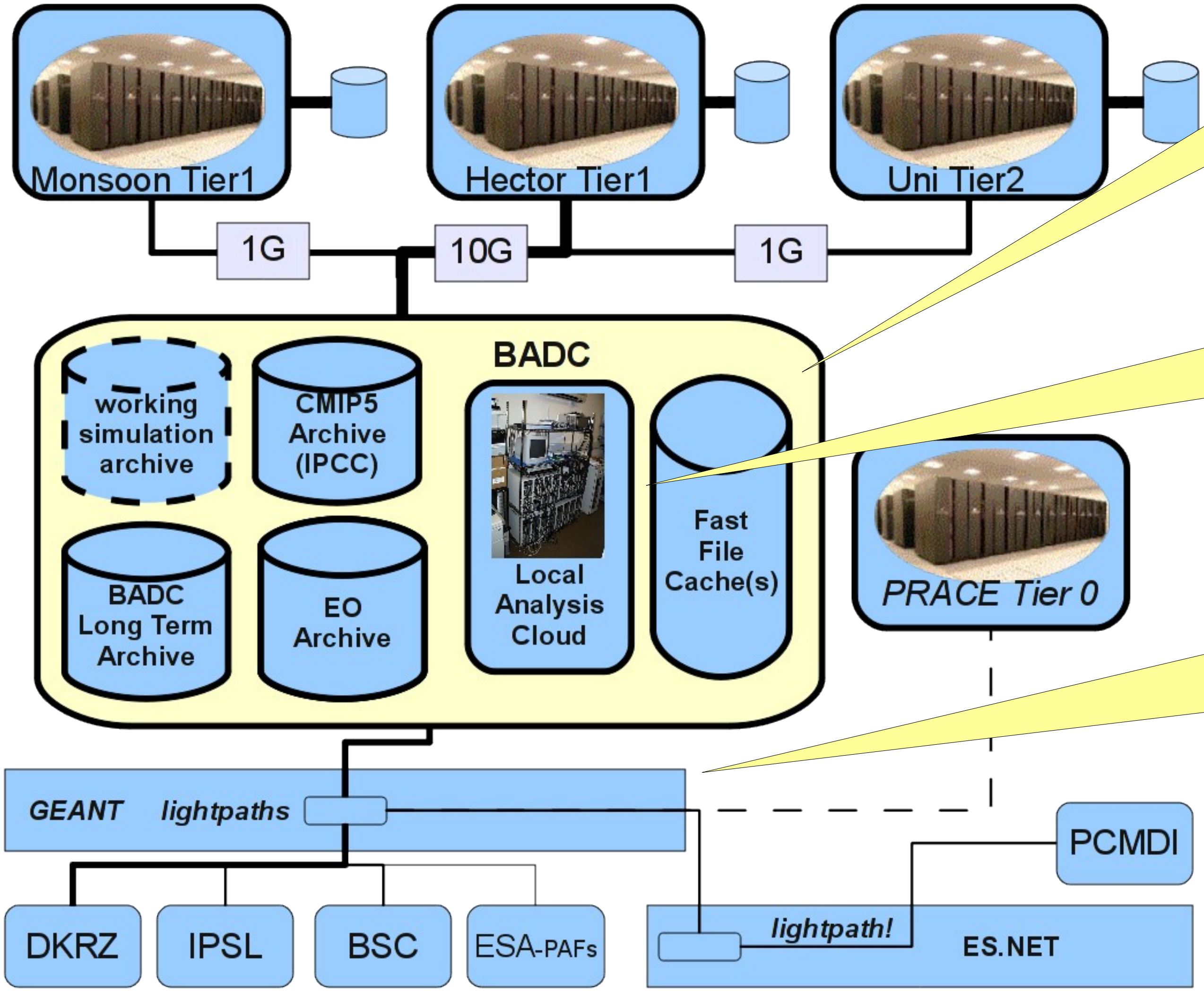
Fifth coupled model intercomparison project (CMIP5) (running now)

- Petabytes of output
- Globally synchronised petascale cache(s)
- Millions of Datasets aimed at different user communities!
- Comprehensive Metadata Structures
- Comprehensive Services

CMIP5 is a GLOBAL problem (the simulations are generated globally and consumed globally)!

Solutions need to be global!

National Problem Too!



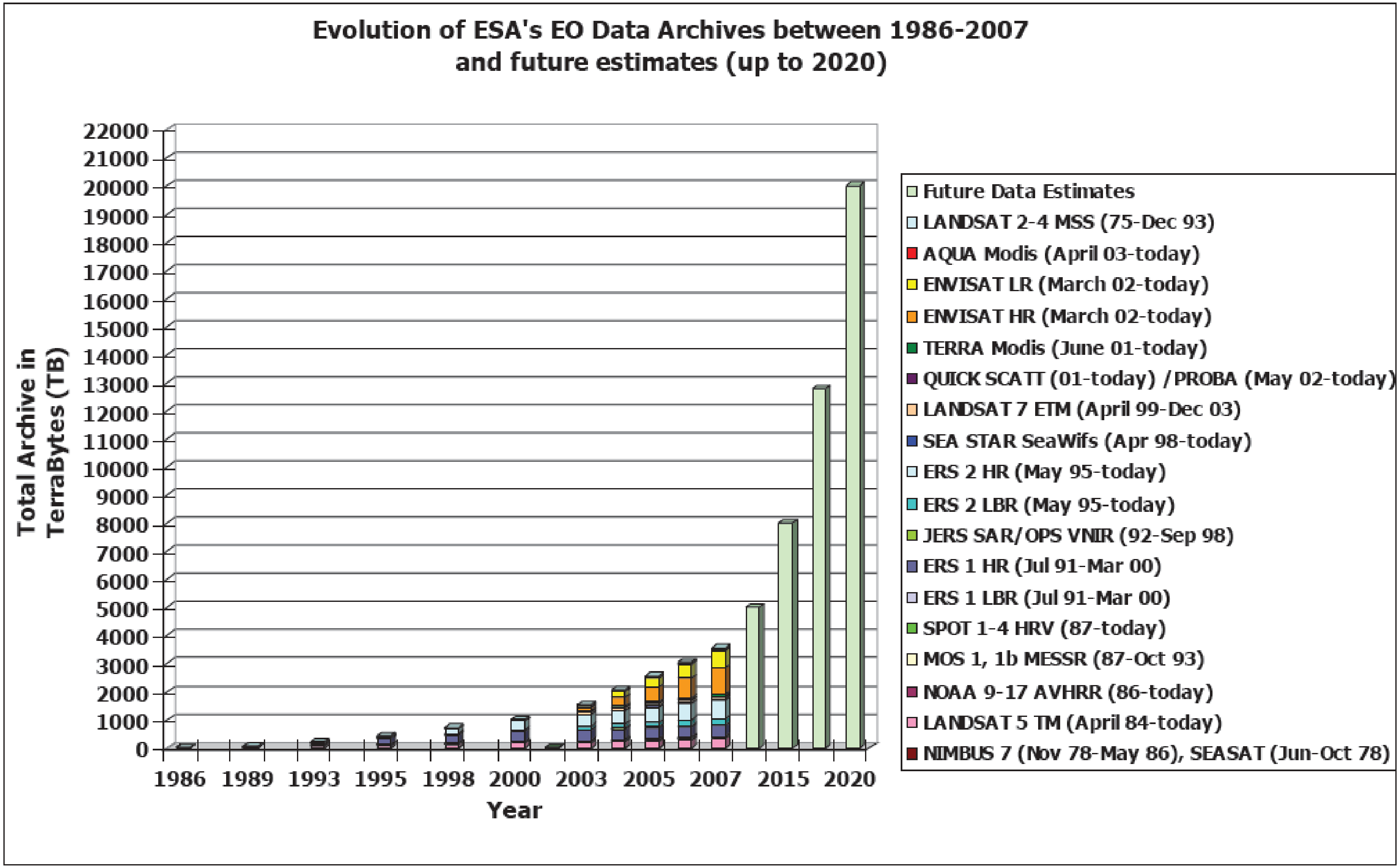
National Caches

Private Cloud (Can't afford network access to commercial cloud)

Optimised Neworks (preferably dynamically allocated light paths)

Planned Evolution

EO Data Deluge

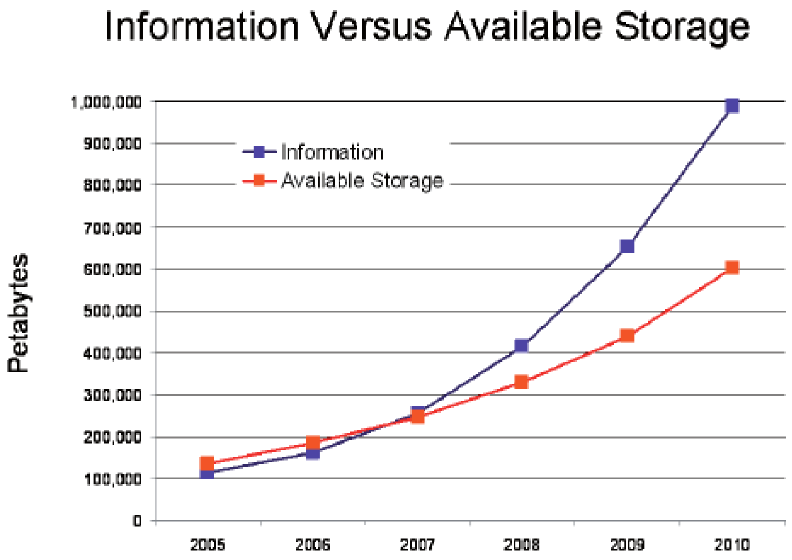


Source: ESA GSCB Workshop June 2009

Observatories and Sensor (networks)

Storage can't keep up!

Figure 2



Source: IDC, 2007

(All data, not just scientific data)

Regardless of how good we are at data systems, science will not escape the general trend: more data being produced than can be stored, which means we need to work smarter:

- Better a priori discrimination of what we should keep
  - Don't even bother writing it to any storage.
- Better documentation of what we have produced, to inform initial decisions about what to keep.
  - Decide quickly about whether to move it to working storage.
- Appraisal of what we have kept (if it's big – don't bother if it's small)
  - Avoid holding data which is irrelevant.

And the Challenge?

Simulation + Earth Observation
+ Sensor Networks
(+looking into the past)

=

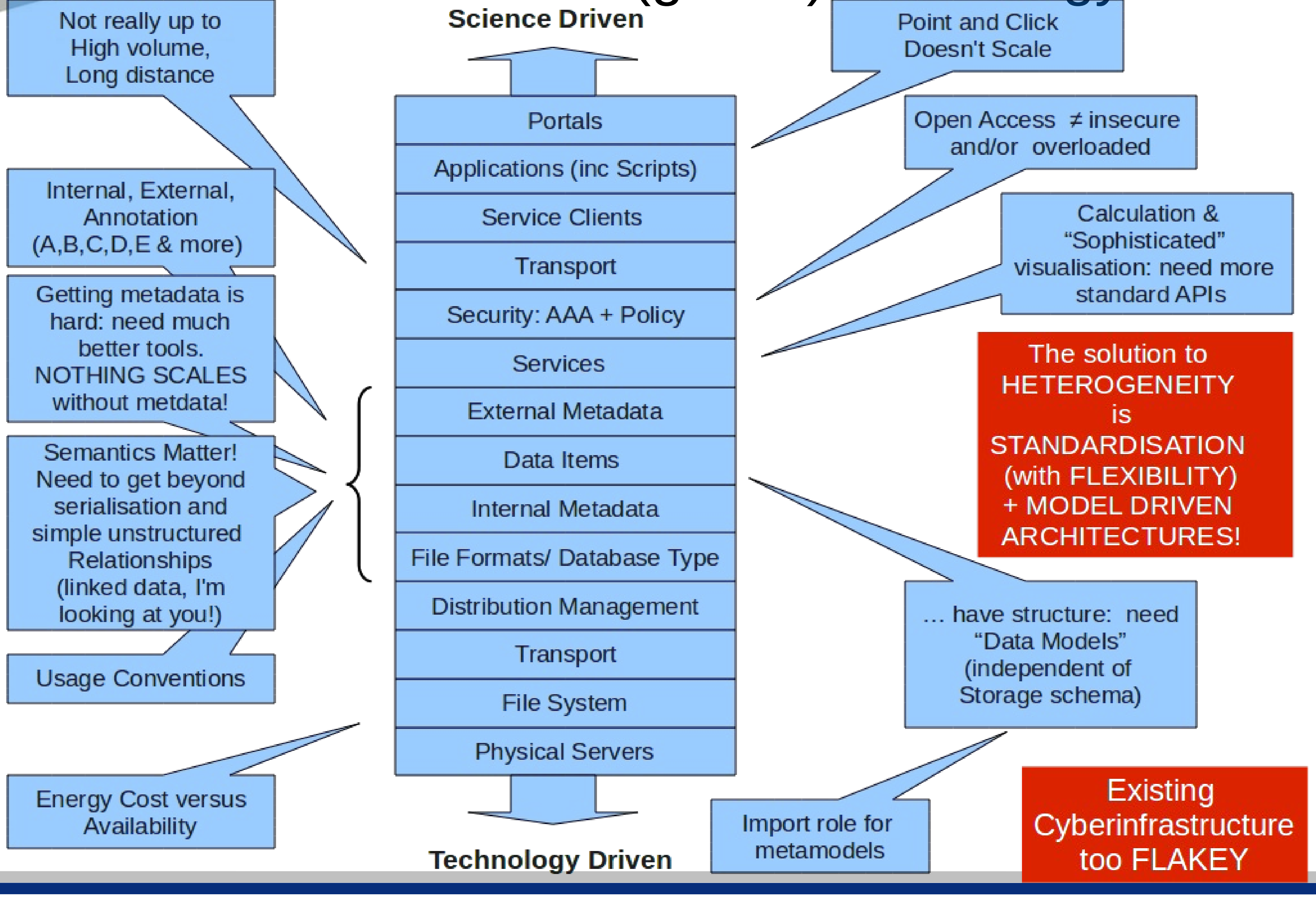
Information about the environment

(all individually increasing their output and proliferating in a heterogeneous and geographically distributed manner)

(which needs integration into a coherent view and interpretation)

Cyberinfrastructure Challenges: from the global large scale data transport and storage, national caches, to automatic/manual metadata creation/entry (*reliable tools to get the metadata to drive it all*) and the systems (including ontology systems) to interpret it all.

And the (global) technology issues?



Science Driven

Portals

Applications (inc Scripts)

Service Clients

Transport

Security: AAA + Policy

Services

External Metadata

Data Items

Internal Metadata

File Formats/ Database Type

Distribution Management

Transport

File System

Physical Servers

Technology Driven

Not really up to High volume, Long distance

Internal, External, Annotation (A,B,C,D,E & more)

Getting metadata is hard: need much better tools. NOTHING SCALES without metdata!

Semantics Matter! Need to get beyond serialisation and simple unstructured Relationships (linked data, I'm looking at you!)

Usage Conventions

Energy Cost versus Availability

Point and Click Doesn't Scale

Open Access ≠ insecure and/or overloaded

Calculation & “Sophisticated” visualisation: need more standard APIs

The solution to HETEROGENEITY is STANDARDISATION (with FLEXIBILITY) + MODEL DRIVEN ARCHITECTURES!

… have structure: need “Data Models” (independent of Storage schema)

Import role for metamodels

Existing Cyberinfrastructure too FLAKEY

Social Challenges

Rewards

Curation

Citation

Licenses & IPR

Trust

Reliance

Plans