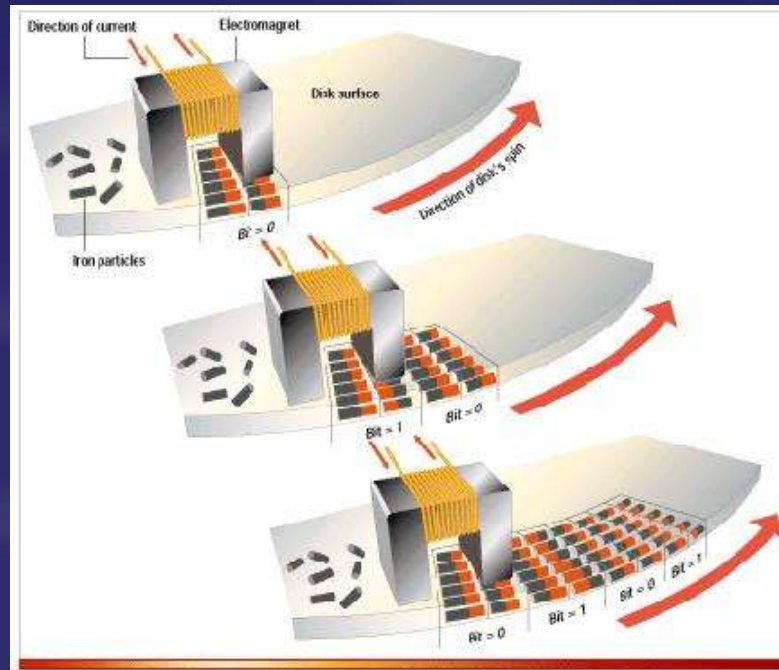# Storage and data services
eIRG Workshop Amsterdam

Dr. ir. A. Osseyran

Managing director SARA

anwar.osseyran@sara.nl

- More data will be produced in 2005 than during the entire existence of humankind!
- By 2005 the amount of information on the planet will increase from 3 million exabytes to more than 40 million exabytes
  - What is the impact of this data explosion?
  - How to deliver, manage, store and recover the data?
  - How to increase the real value of the data?

Scientific Experiments Pull

- Scientific experiments start to generate *lots* of data
    - Bio-informatics queries: **500-1000 GByte** in dbases
    - medical imaging (fMRI): **~ 1 GByte** per measurement
    - Satellite world imagery: **~ 5 TByte**/year
    - Current particle physics: **1 PByte** per year
    - LOFAR (2007): **>25 PByte** per year
    - LHC physics (2007): **10-30 PByte** per year
- Data is often very distributed and dynamic according to the acquisition situation
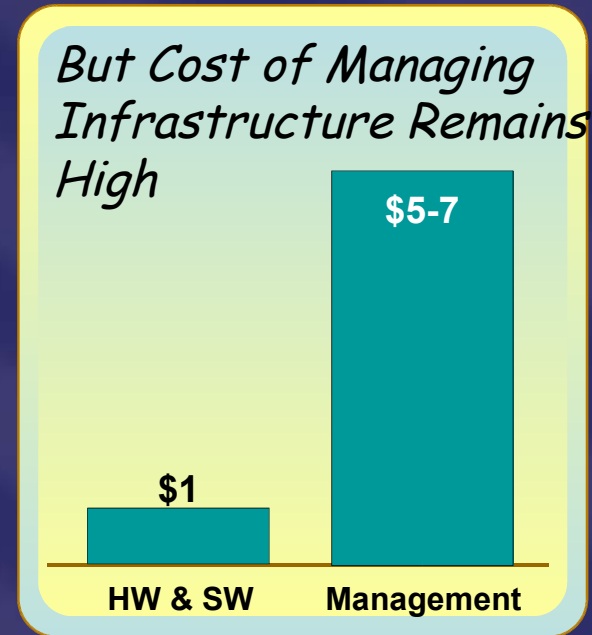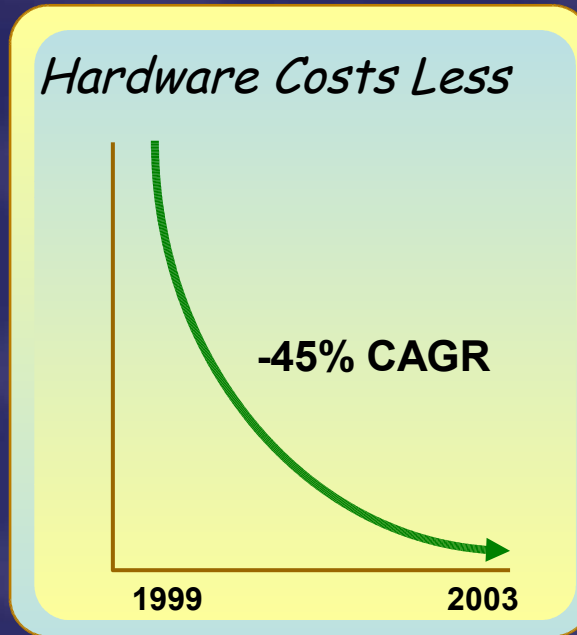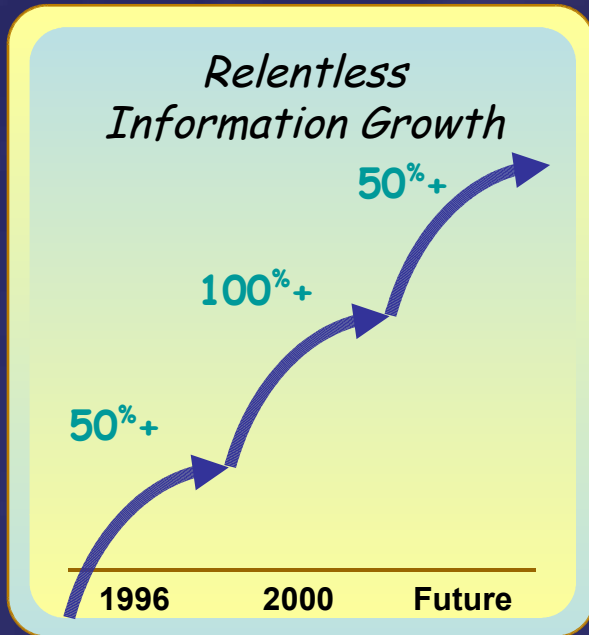
- Processing power doubles every   18 month
- Memory size doubles every   12 month
- **Network speed doubles every   9 month**
- Something has to be done to harness this development
  - Virtualization of ICT resources
    - ‣ Internet
    - ‣ WEB
    - ‣ **Grid**
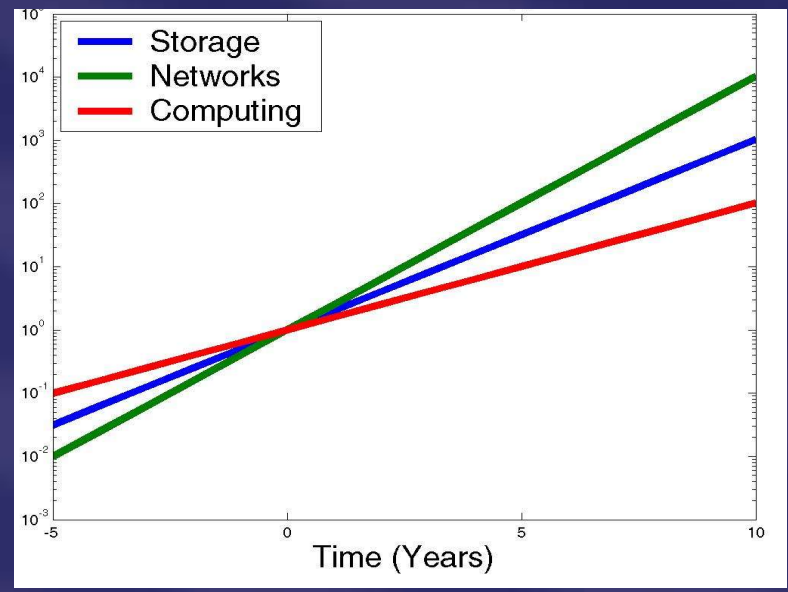
Managing Costs

- Increasing information needs of mission critical data
- The Manpower required to manage storage is high
- Management of storage per GB is at least five fold higher than price
- Management of future capacity needs

Relentless Information Growth

50%+

100%+

50%+

1996   2000   Future

Hardware Costs Less

-45% CAGR

1999   2003

But Cost of Managing Infrastructure Remains High

$5-7

$1

HW & SW   Management

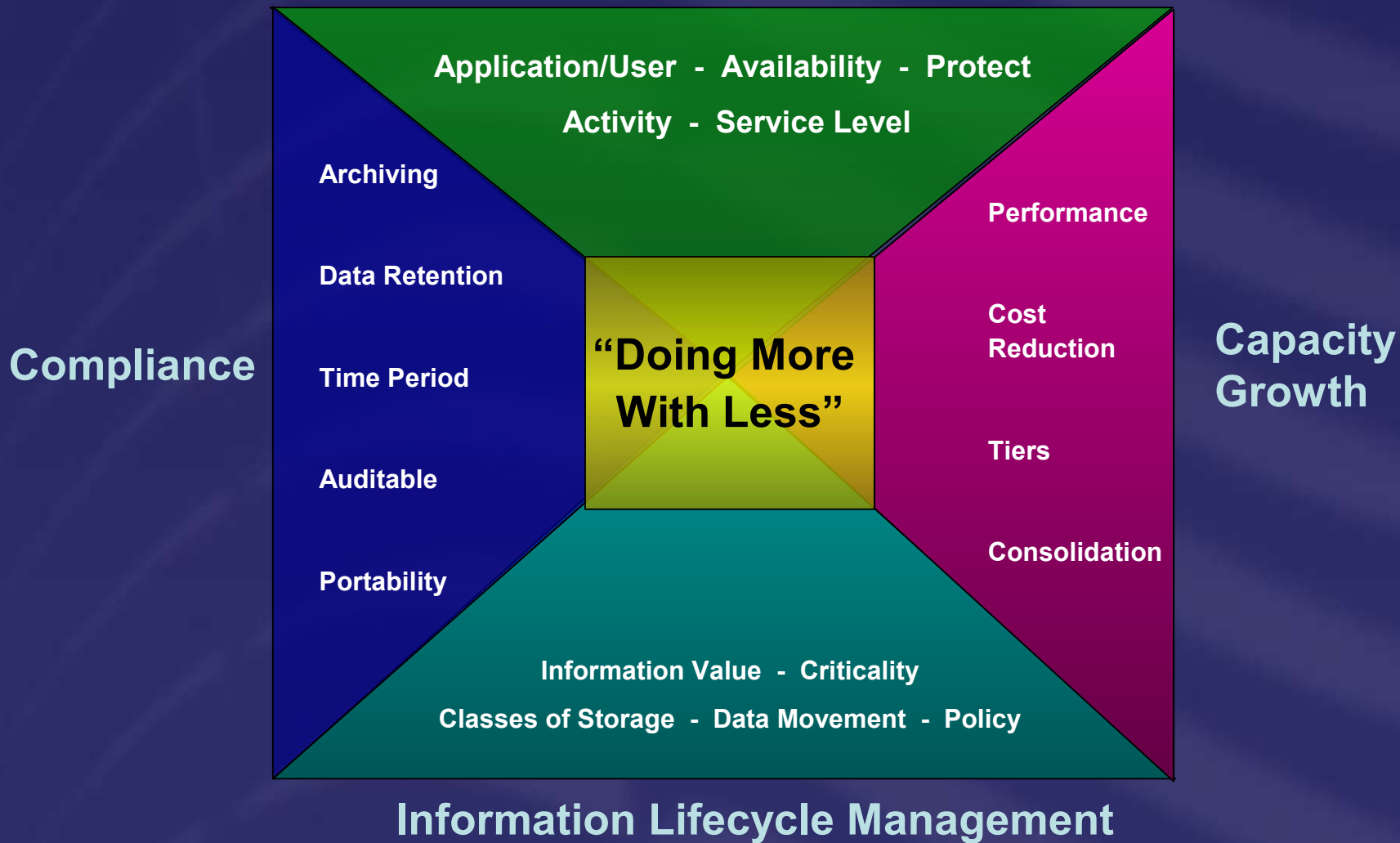Source: IDC

What our End Users are asking for?

- Transparent access to distributed data archives
- Secure Backup and Archiving
- Storage interoperability regardless of platform
- Better configuration, monitoring and management tools
- Distribution of excess load throughout the WAN
- Reduced downtime
- Prediction of future storage needs from current usage patterns
- Alignment with business operations
- Service levels & Cost management
- Management of heterogeneous environment from a single interface
- Storage resource management and associated storage services, e.g. capacity planning, data migration, performance planning, etc.
- Improved TCO – with high availability, high volume processing and reduced overall cost per transaction
- On-demand services

Infrastructure Complexity & Legislations

- Infrastructure Complexity: Today most organisations have heterogeneous environments...

- Creating Challenges around;
  - Scalability, Reliability, Security
  - Disaster recovery, Business continuity
  - Host, Windows/UNIX clients, Linux too
  - Legacy applications

- Regulatory compliance heavily impacts management of information life cycle

- Creating Challenges around:
  - Managing the data over its lifetime -> ILM
  - Regulation compliance is never ending !!!
  - Grading of data is required
  - Means more spending

Today's Infrastructure needs high availability (24/7), scalability and robust infrastructure

Today's Storage Challenges

Backup and Recovery

Compliance

Capacity Growth

Application/User - Availability - Protect

Activity - Service Level

Archiving

Data Retention

Time Period

Auditable

Portability

Performance

Cost Reduction

Tiers

Consolidation

"Doing More With Less"

Information Value - Criticality

Classes of Storage - Data Movement - Policy

Information Lifecycle Management

- Can the research community in Europe benefit from having a shared approach to the increasing storage needs?
- Which strategies should we develop to make our IT spending keep pace with storage capacity needs?
- What is the most cost-effective way to scale - Scaling up or scaling out? What about data ownership and storage costs?
- Should we establish a distributed shared network of a (limited) number of (large) storage facilities?
- If yes, should that be tightly or loosely coupled ? Project-based or centrally co-ordinated?
- How should the responsibility for data integrity and ultimate backup and archiving be then organized?
- Does this yield the redundancy required for advanced data recovery?
- What should be preferred funding and accounting model?

- Should this approach be just limited to large centralized data generators (like large equipment or large supercomputers)?
- Should we relate storage-tiering models with processing-tiering models and where could efforts be combined?
- To which extent should a European data grid infrastructure be generic or application/community specific?
- Should there be European guidelines for:
    - e-infrastructures
    - data classification
    - storage site classification
    - vendor compliance
    - reliability, confidentiality, availability, integrity, accessibility and performance
    - certification and auditing
- Anything Else?