Datensicherung im Wandel

Dr. Thomas Eifert

IT Center | RWTH AACHEN UNIVERSITY

RWTH Aachen University

Eine der TU9-Hochschulen in Deutschland

45.377 Studierende (davon 9.651 internationale Studierende aus 125 Ländern)

547 Professoren, 5564 Wissenschaftliche Mitarbeiterinnen und Mitarbeiter

Unter den Top 3 der deutschen Hochschulen im Drittmittelaufkommen

Natur- und Ingenieurwissenschaften mit sehr großen Mengen an Forschungsdaten
- Simulations- und Simulationsbasierte Wissenschaften
- Experimentell arbeitende Wissenschaften

RWTH aus der Perspektive des zentralen Backup-/Restore-Service

100%-Abdeckung der Einrichtungen als Nutzende des zentralen Backup/Restore Service

Aktuelle Infrastruktur für TSM (seit Juli 2015)

- 4 Server, jew. 1 TB RAM, 2*40Gbit/s
- 1 PB Disks
- TS 3500 Library
- 48 x IBM TS1150 tape drives

Seit 2010: Service Level Agreement mit Einrichtungen:
„any restore at any time with client‘s wirespeed“

presented on TSM Symposium 2011 and in PIK publication, Vol. 35(3), p. 195–198, DOI: 10.1515/pik-2012-0032

Service für Uni Paderborn und FH Aachen

Über mich

- Verantwortlich für Aufbau des zentralen Backup-Service ab 2000
- Selfcare-Portal für TSM-basierten Service: 2007
- SLA für Restore: 2011
- Umbenennung des Service: „Restore-Service“: 2011

- Rollenwechsel 2013: CTO

  Verantwortlichkeiten
  - Technologische Strategie
  - Fördermittel

Datensicherung ?

Definition:

Durch technisches Versagen, versehentliches Löschen oder durch Manipulation können gespeicherte Daten unbrauchbar werden bzw. verloren gehen. Eine Datensicherung soll gewährleisten, dass durch einen redundanten Datenbestand der IT-Betrieb kurzfristig wiederaufgenommen werden kann, wenn Teile des operativen Datenbestandes verloren gehen.

aus: BSI IT Grundschutz

Kürzer:

Schutz von Datenbeständen gegen Verlust

Also: „Es kommt nichts weg“

➔ Ziel klar !

Wege zum Ziel ⇔ Ausgangspunkte

IT Center | RWTH AACHEN UNIVERSITY

„Ziel klar“ ➔ Dienstgüte-Zusage

- Forschung und Lehre abhängig (!) von
  - Datenspeicherung
  - Zugriff

  - Permanent
  - zuverlässig

- Wissenschaftler müssen sich auf zuverlässige Datenhaltung verlassen können (andernfalls: Nebenwege ➔ konterkariert strukturierte Speicherung und FDM)

- Sichere Datenhaltung als Grundbestandteil der weiteren Digitalisierung

„Datensicherung“

- Datensicherung: Gesamtheit der Maßnahmen zum Schutz gegen Datenverlust
- Backup: (Blindes) Abschreiben eines Datenbestandes auf separate Plattform an separatem Standort
- Tape: Speichermedium

Datenhaltung: Was haben wir?

- Online-Storage („General purpose“)
  - Instituts-Fileserver
  - Lokale Platte auf Endgerät
- Online-Storage
  - HPC

    Singuläre Größen
  - Kollaborations-Plattformen
    - Web-Server, Wikis, …
- Nearline
  - Archiv

➔ Speicherorte wertvoller Daten ➔ vor Verlust zu sichern

Was haben wir noch:

- Transaktions-Speicher („strukturierte Daten“)
  - Datenbanken
    - Verwaltung
      - ERP (~100 TB)
      - Campus Management (~0.6 TB)
    - IT-Administration
      - IdM
      - NOC
      - …
    - Datenbanken für Nutzer-Dienste
    - FD-Metadaten
  - Sharepoint

Benötigt Sicherung auf Anwendungsebene

Datensicherung: Schutz von Datenbeständen gegen Verlust

Gegen Fehlbedienung / Software-Fehler auf Anwendungsebene

- Kopieren ➔ Backup
- Versionierung, ggf. mit Überschreib- / Lösch-Schutz
- Verzögertes Ausführen von Löschungen

Datensicherung: Schutz von Datenbeständen gegen Verlust

Gegen Ausfall des Speichersystems /-Mediums (Hardware, Gebäude, ..)

- Kopieren
  - Replikation (Cloud),
  - Backup

  - Bei großen Speichersystemen: Restore kaum in sinnvoller Zeit möglich

- Verteilen ➔ Redundanz, EC

Foto: © Ralf Röger

Datensicherung: Schutz von Datenbeständen gegen Verlust

Gegen Software-Fehler auf Speichersystem-Ebene

➔ Kopieren mit Plattform-Wechsel

➔ Backup

Laut einer Studie an einer englischen Universität ist es egal, in welcher Reihenfolge die Buchstaben in einem Wort sind.

Luat enier sidtue an eienr elgnhcsien uvrsnäiett, ist es eagl in wcheler rhnfgeeloie die bstuchbaen in eniem wrot snid.

Kosten von Speichermedien

Quantitative Abschätzung:

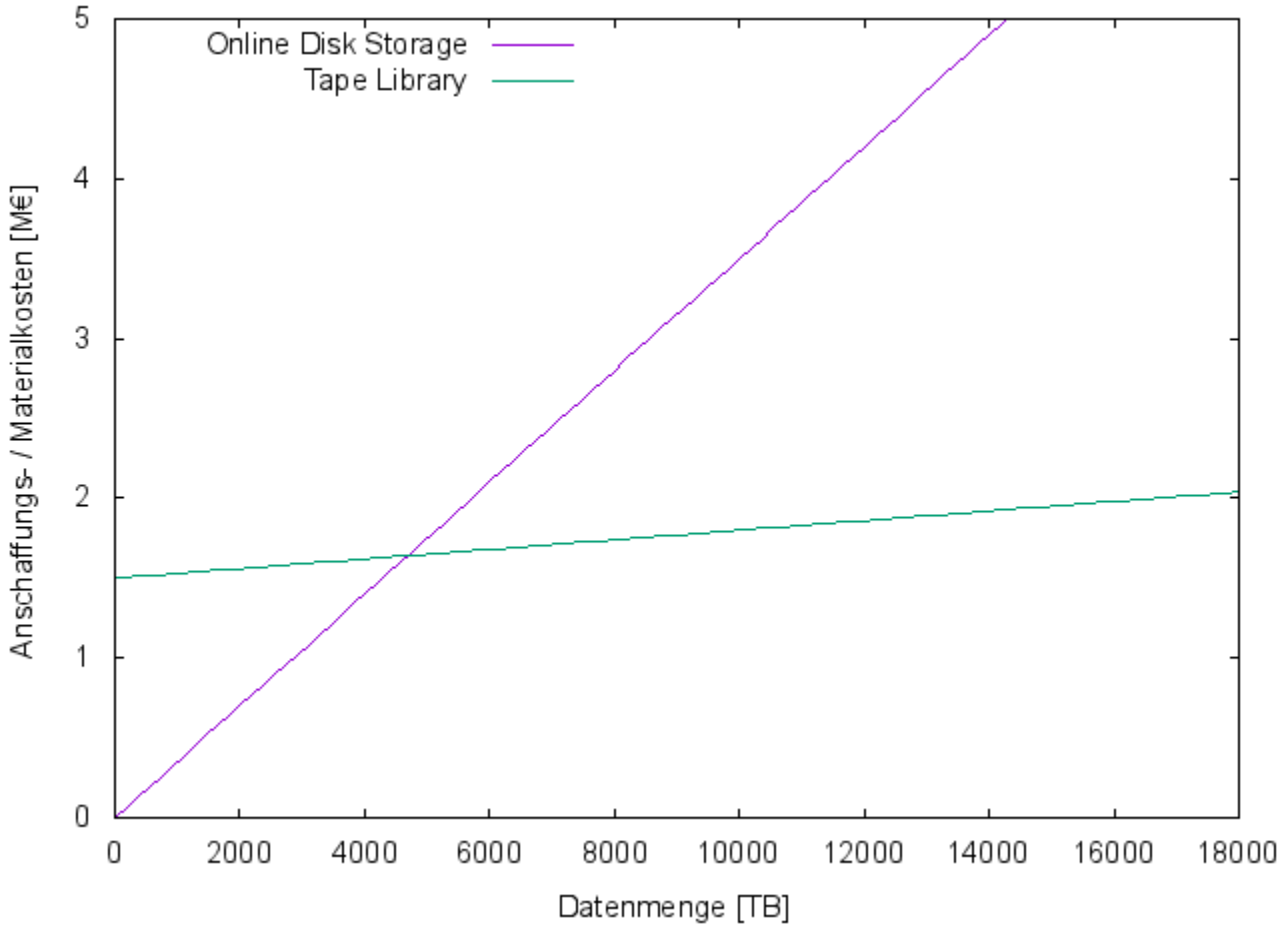
- Alle Datenbestände zeigen steile Wachstumskurve ➔ Kosten
  - vervielfacht bei Redundanz

- Aktuelle Grenzkosten:
  - Disk: ~35 ct / GB ➔ 350 k€ / PB, ~2..3 kW / PB
  - Tape: ~3 ct / GB ➔ 30 k€ / PB
  - AWS Glacier: ~0.5 ct / (GB*M), nach 60 Monaten ~30 ct / GB zzgl. Abruf-Kosten
  - AWS S3 „Infrequent Access“: ~1.5 ct / (GB*M) ➔ nach 60 Monaten ~90 ct / GB

Achtung: keine Vollkosten!

➔ 4 PB: 1.4M€ (Disk) bzw. 1.6M€ (Tape)
20 PB: 7 M€ (Disk) bzw. 2.1M€ (Tape)

Disk / Tape – Szenarien



Annahmen

- „Festplatten sind online“
  ➔ Plattenbasierte Speicherung erlaubt interaktive Nutzung
  - Verwenden von „Nearline-Platten“? (Strikt auf preisgünstige Kapazität optimiert, ca. 15 ct/GB)

- „Tape ist Kopie“
  - Grundannahme bei Backup: Das Original liegt räumlich an anderer Stelle
  - Archiv: Primärspeicher ➔ Annahme gilt nicht!

  - Benötigen wir 2* Tape-Infrastruktur, verteilt auf 2 Gebäude?

➔ Der Vergleich ist sicher komplizierter!
  ➔ Funktionale Eigenschaften (Zugriff (→Disk), ruhende Daten (→Tape), ..) hinsichtlich Data Access Pattern zu bewerten

  ➔ Abschätzung Betriebsaufwand / TCO: Robotik vs. ~x000 Festplatten

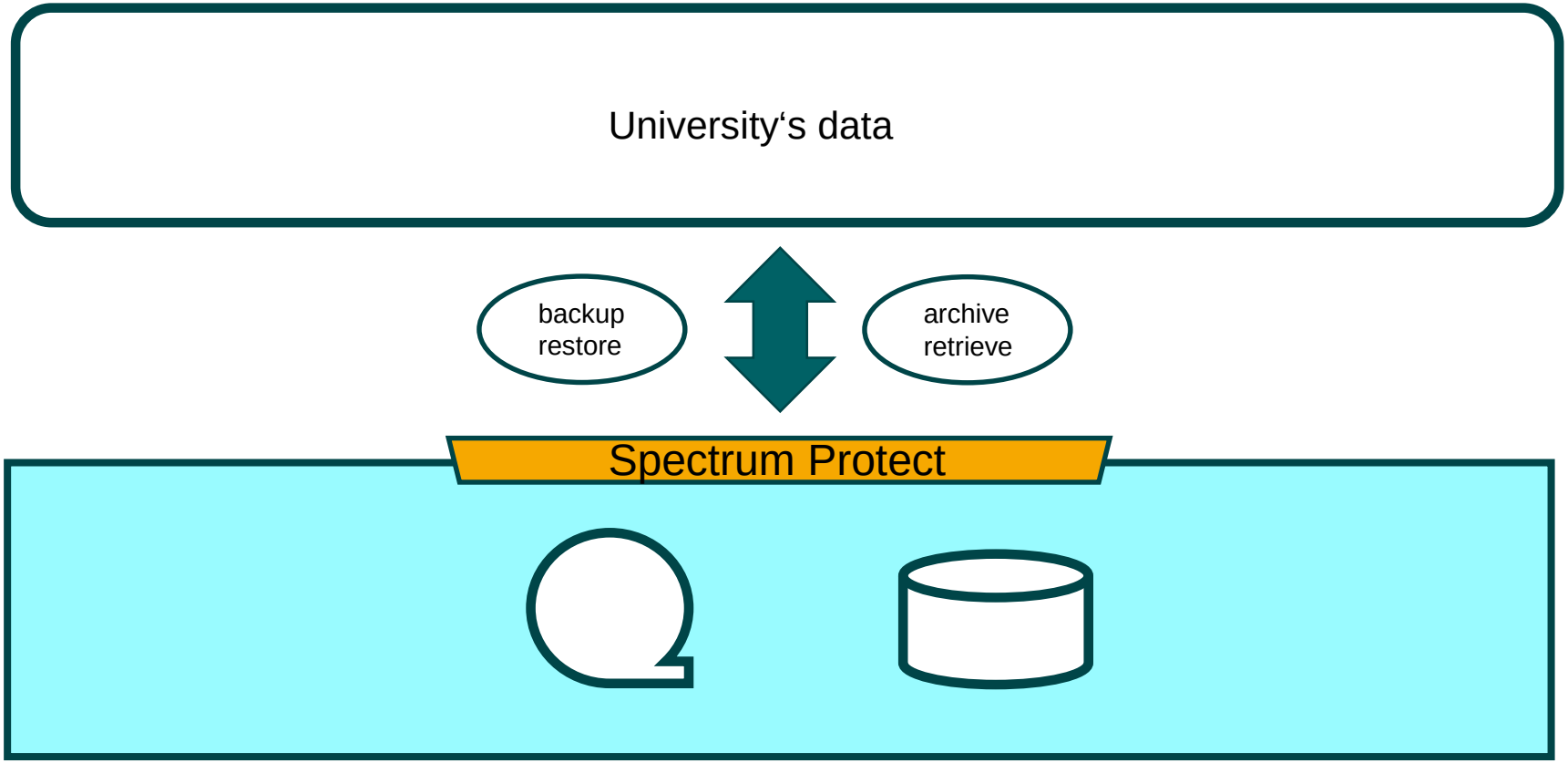
  ➔ Eingesetzte Technologie(en) abh. von Datenvolumen und Gesamtstrategie

Zurück zu Sicherungsszenarien

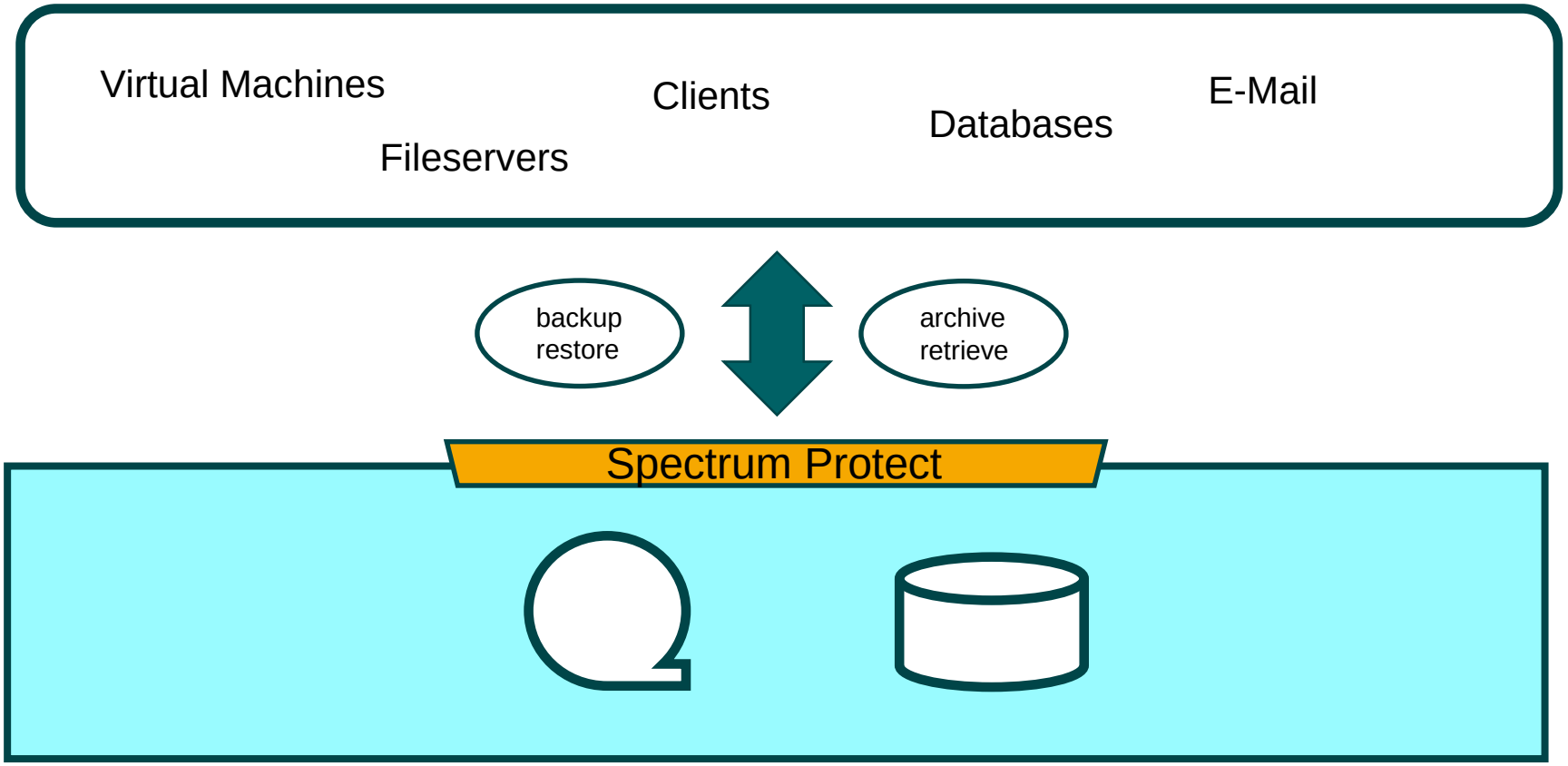
- Was wird wo gespeichert?
- Wie gesichert?

- Und wie könnte das künftig aussehen?

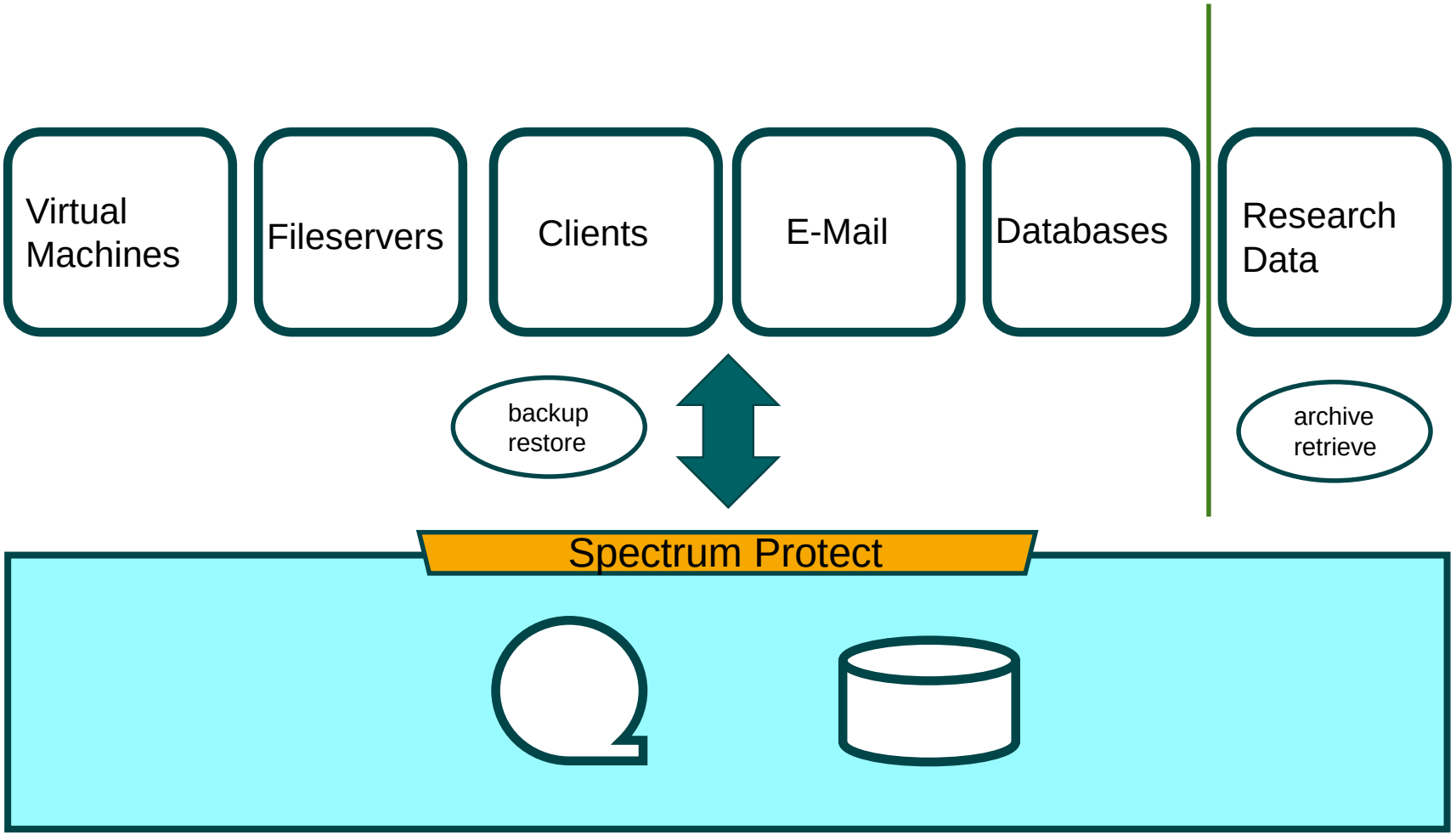
Backup aller Daten



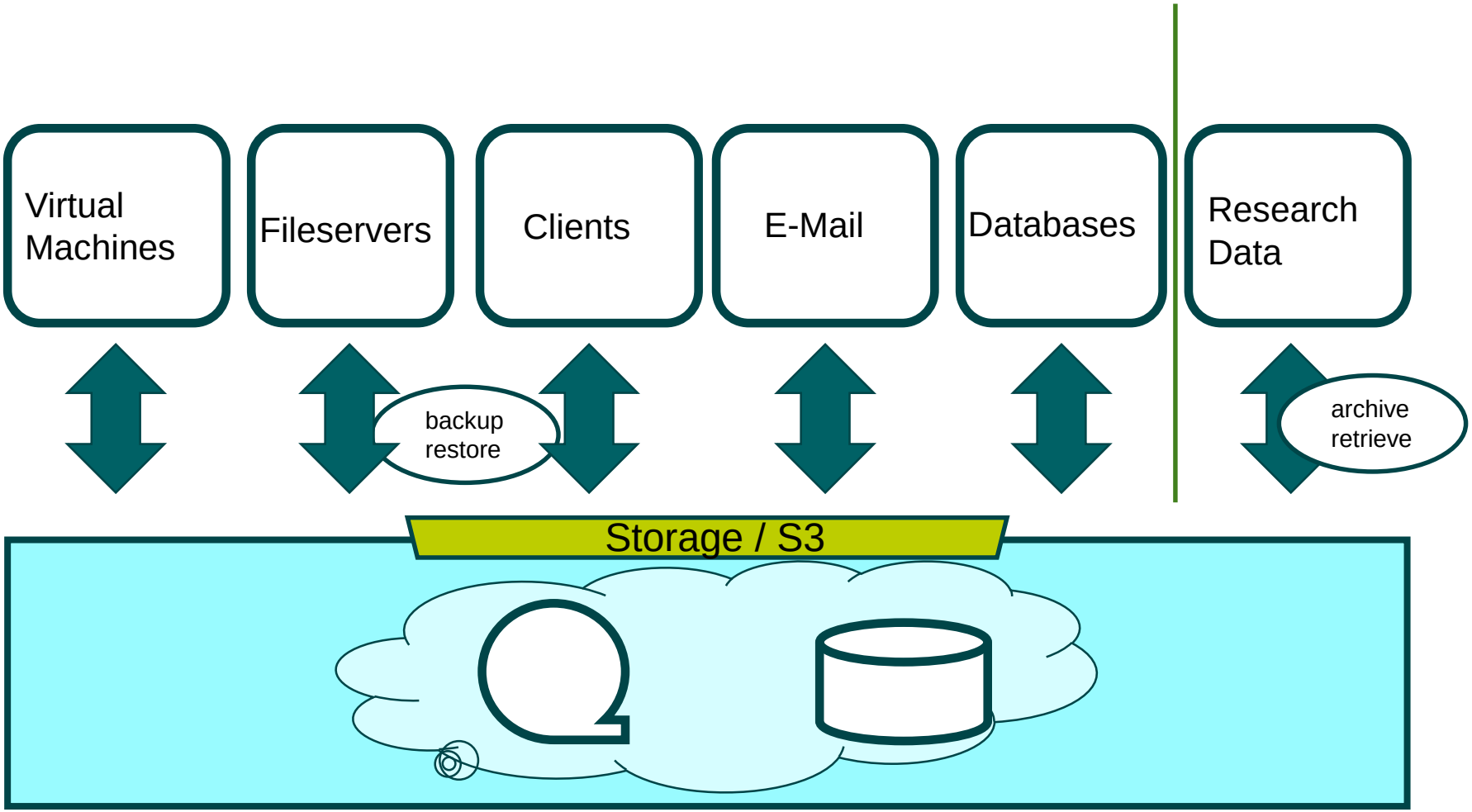
Backup: Differenzierte Datenquellen



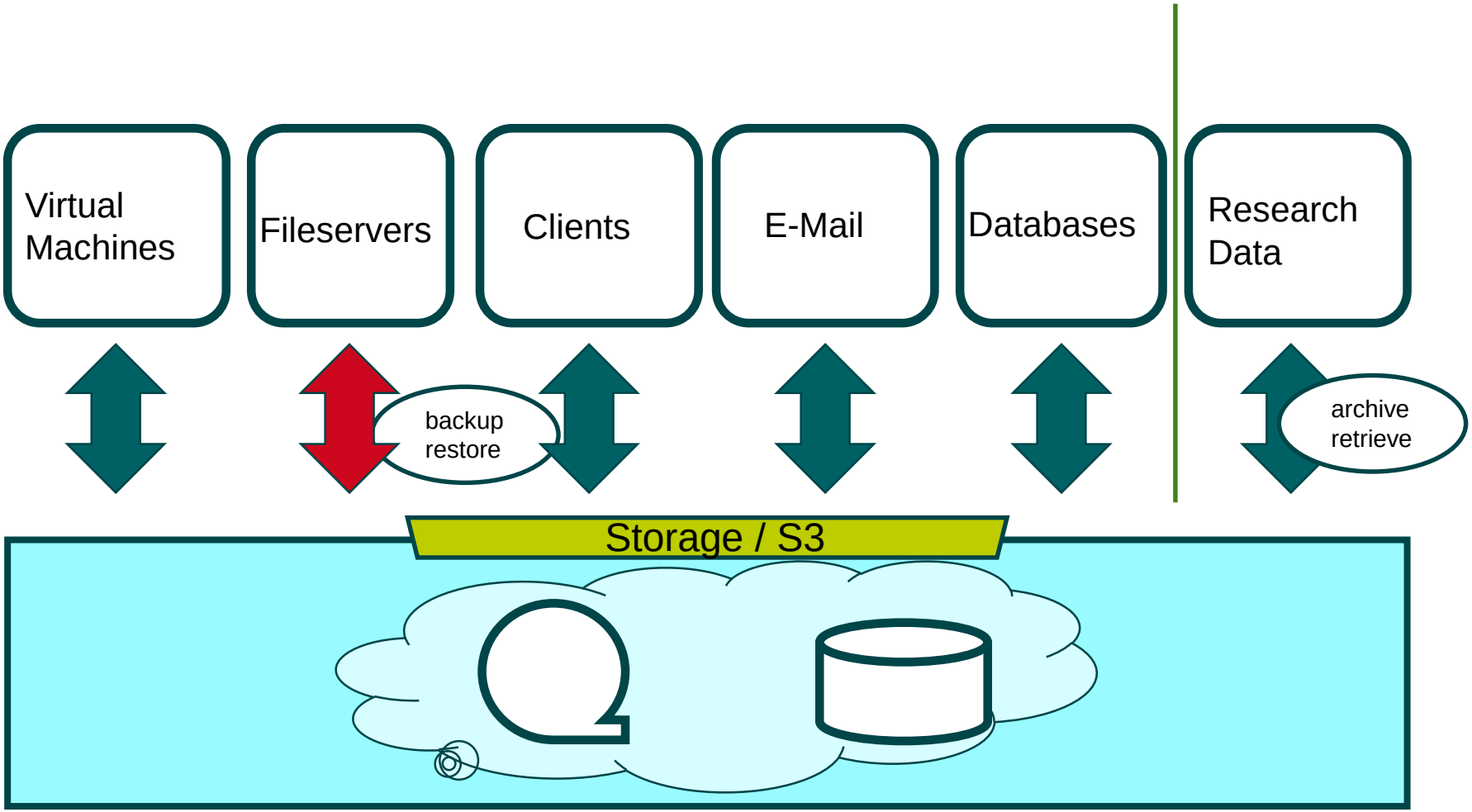
Backup: Einer für alle?



Backup: .. Oder für jeden das Beste?



Dienste-spezifische Sicherungsstrategie



Änderung der Datenhaltung durch FDM

Bisher: Fileserver, dezentral organisierte Datenspeicherung

Neu:

- Aufteilung der Aufgaben: „normaler“ Fileserver und FD-Speicher
- „normal“
  - Dezentral organisiert

- FD: zentrale Organisationselemente
  - IdM
  - Organisatorische Metadaten
    - Projekt-Zuordnung
    - Speicher-Zeiten
  - Service:
    - Metadaten-Speicher und –Werkzeuge
    - Integration mit anderen FD-spezifischen Services

Speicherung / Sicherung von Daten - heute

Typ der Daten	Charakteristik Datenzugriff	Primärer Speicher	Sicherung	
Strukturierte Daten / DB	In use	File / Block on Disk	Backup to Tape	
Unstrukturierte Daten / Files	In use	File on Disk	Backup to Tape	
Archiv	Low use / at rest	Tape	Spiegel FZJ	

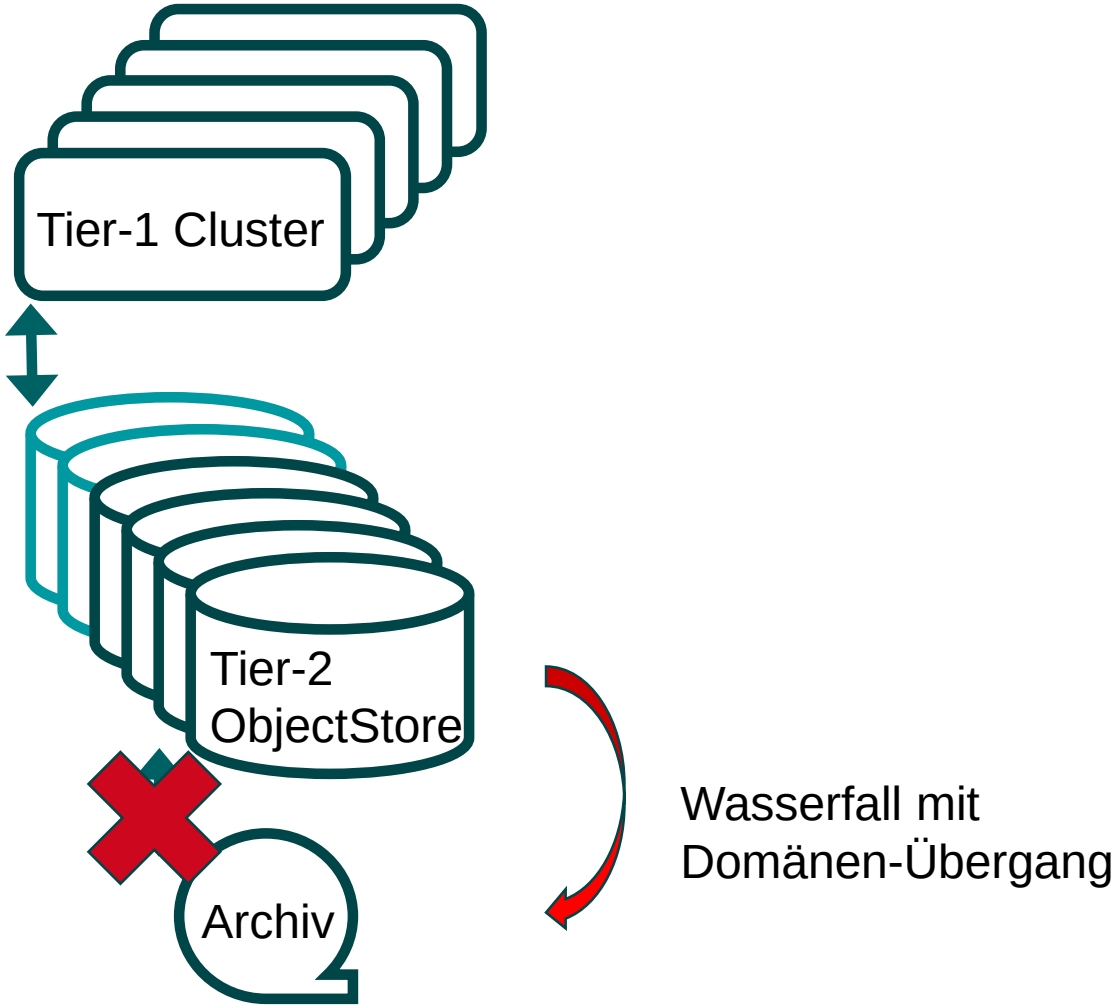
Speicherung / Sicherung von Daten – künftig

Typ der Daten	Charakteristik Datenzugriff	Primärer Speicher	Sicherung	Expansion und Sicherung
Strukturierte Daten / DB	In use	File / block on Disk	Backup to Object	
Unstrukturierte Daten (allg.)	In use	File on Disk / Cloud	➔ s. nächste Seite	
Unstrukturierte Daten (FD)	In use	Object on Disk	Georedundanz, EC / Replikation	2nd layer Object
Archiv	Low use / at rest	Object on Disk	Georedundanz, EC / Replikation	2nd layer Object

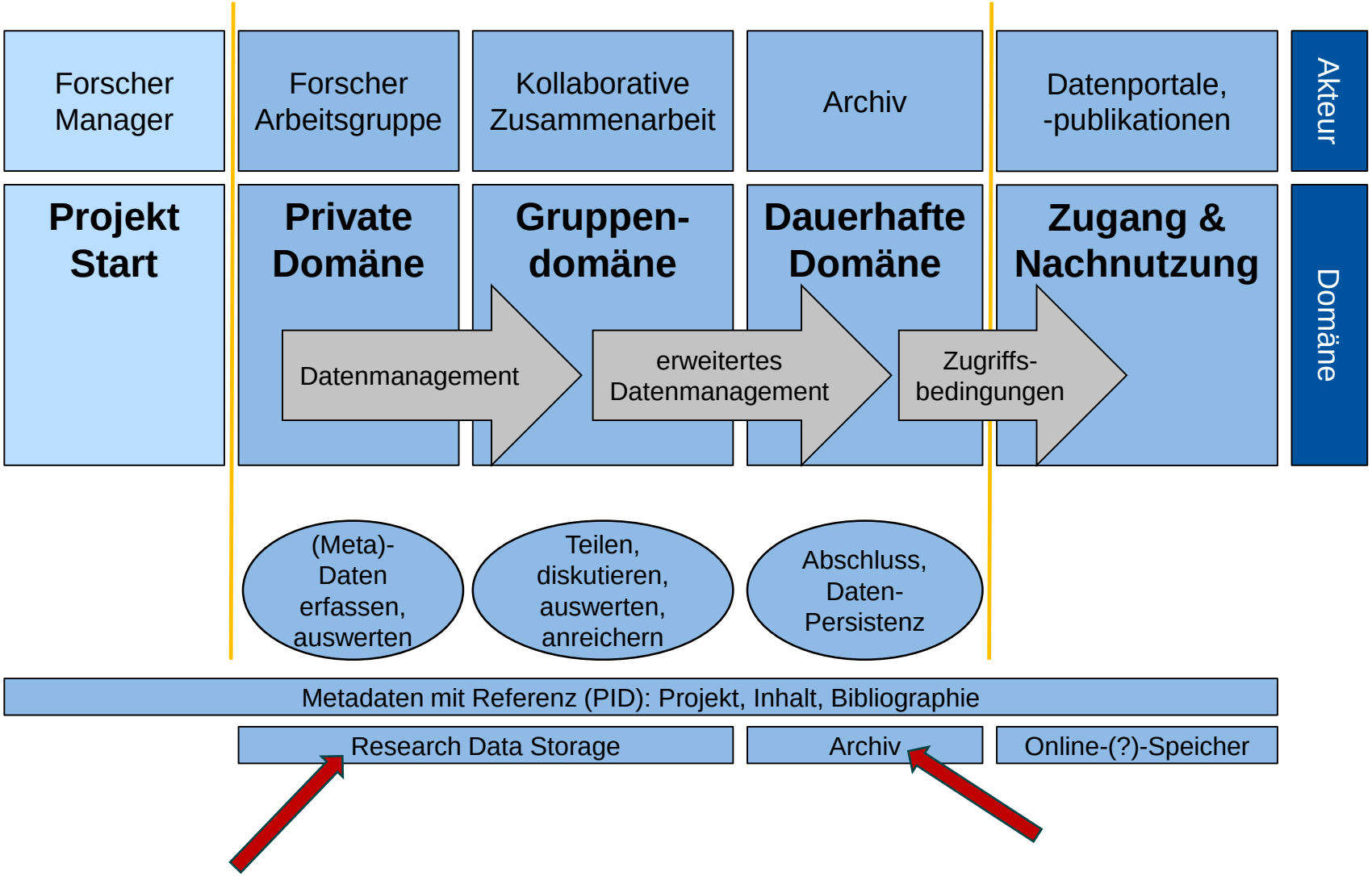
„Datensicherung“

- Datensicherung: Gesamtheit der Maßnahmen zum Schutz gegen Datenverlust
- Backup: (Blindes) Abschreiben eines Datenbestandes auf separate Plattform an separatem Standort
- Tape: Speichermedium
- Archiv: Bewusstes Ablegen eines (abgeschlossenen) Datenbestandes auf zuverlässigem Langzeitspeicher

FD-Storage: 3-Tier-Speicherhierarchie (Wasserfall-Modell)



Erweitertes Domänenmodell: Forschungsdaten-Lebenszyklus



Sicht auf Daten

- Traditionelle Sicht auf Daten:
  - Die wertvollsten Daten sind die aktuellen
  - Die Relevanz nimmt mit der Zeit ab
  - Annahme (z.B. ISP-Symposium):
    archivierte Daten müssen verwahrt werden, in der Praxis wird nie mehr drauf zugegriffen

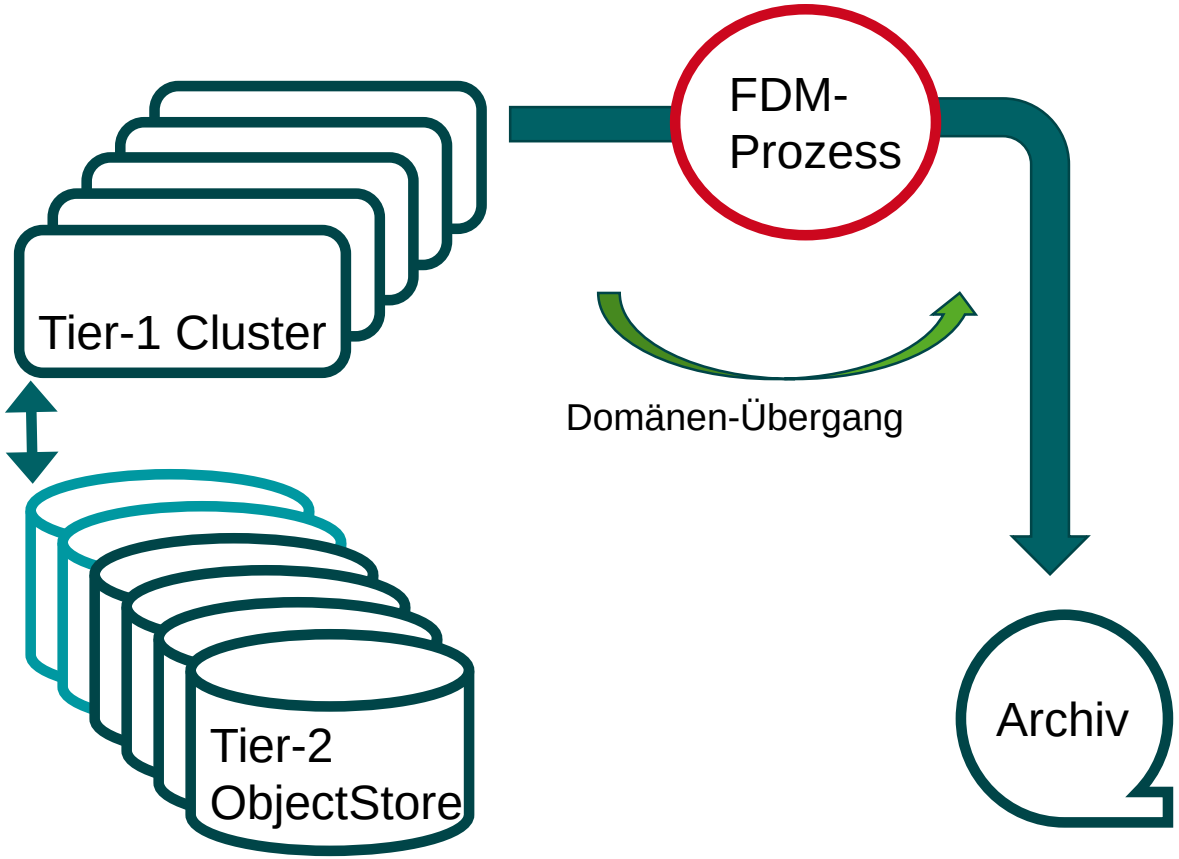
  ➔ Speicher-Hierarchie getriebenes Wasserfall-Modell

- FD-Sicht auf Daten:
  - Es werden viele Experimente gemacht
  - Im Zuge der Auswertung wird sortiert und angereichert
  - Daten, die zu Erkenntnissen führen oder Thesen untermauern, sind die wertvollsten und müssen archiviert werden.

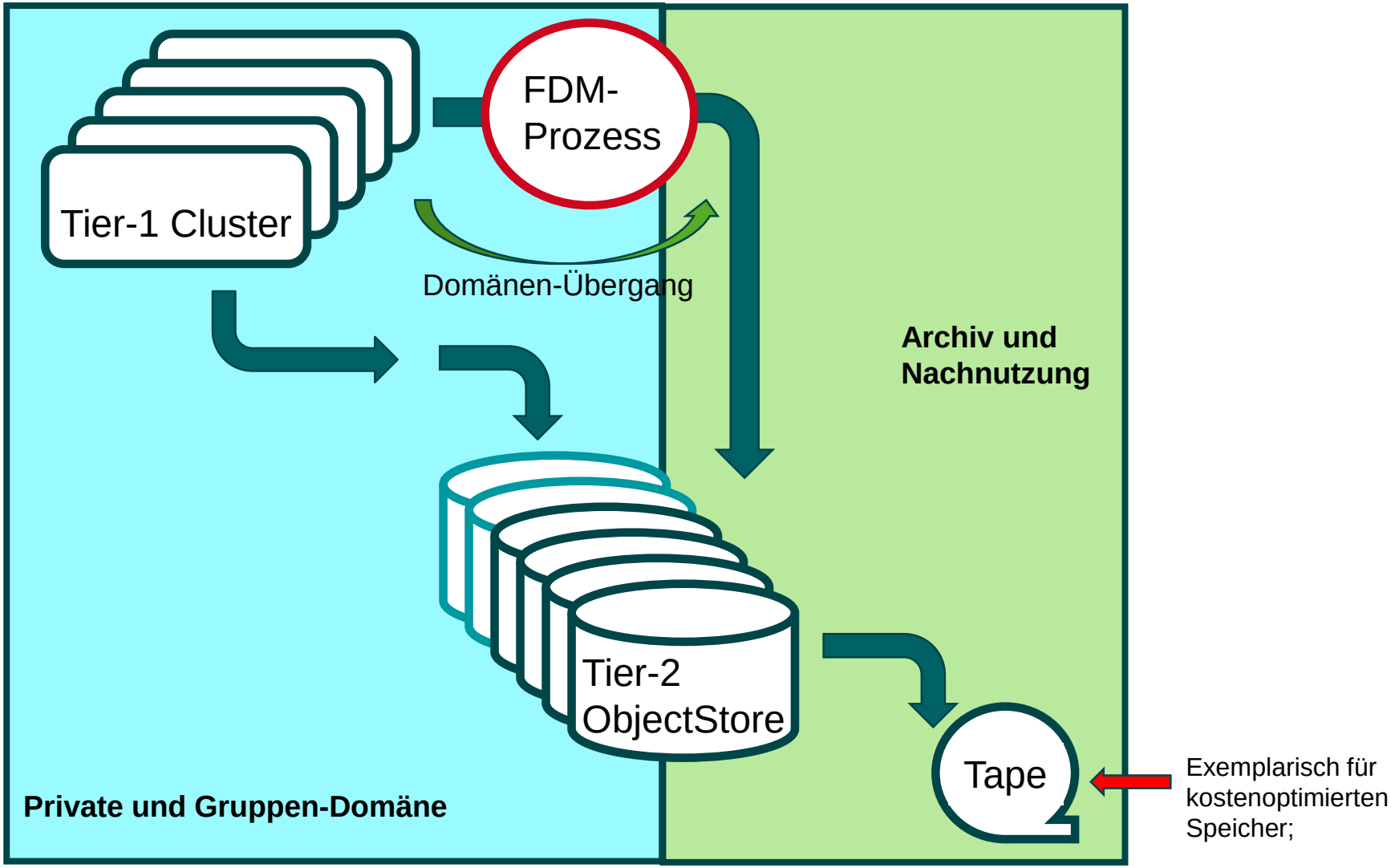
  ➔ Das Archiv enthält das „Daten-Gold“

➔ Widerspruch

FD-Storage: 3-Tier-Speicherhierarchie (Archiv auf Tape)



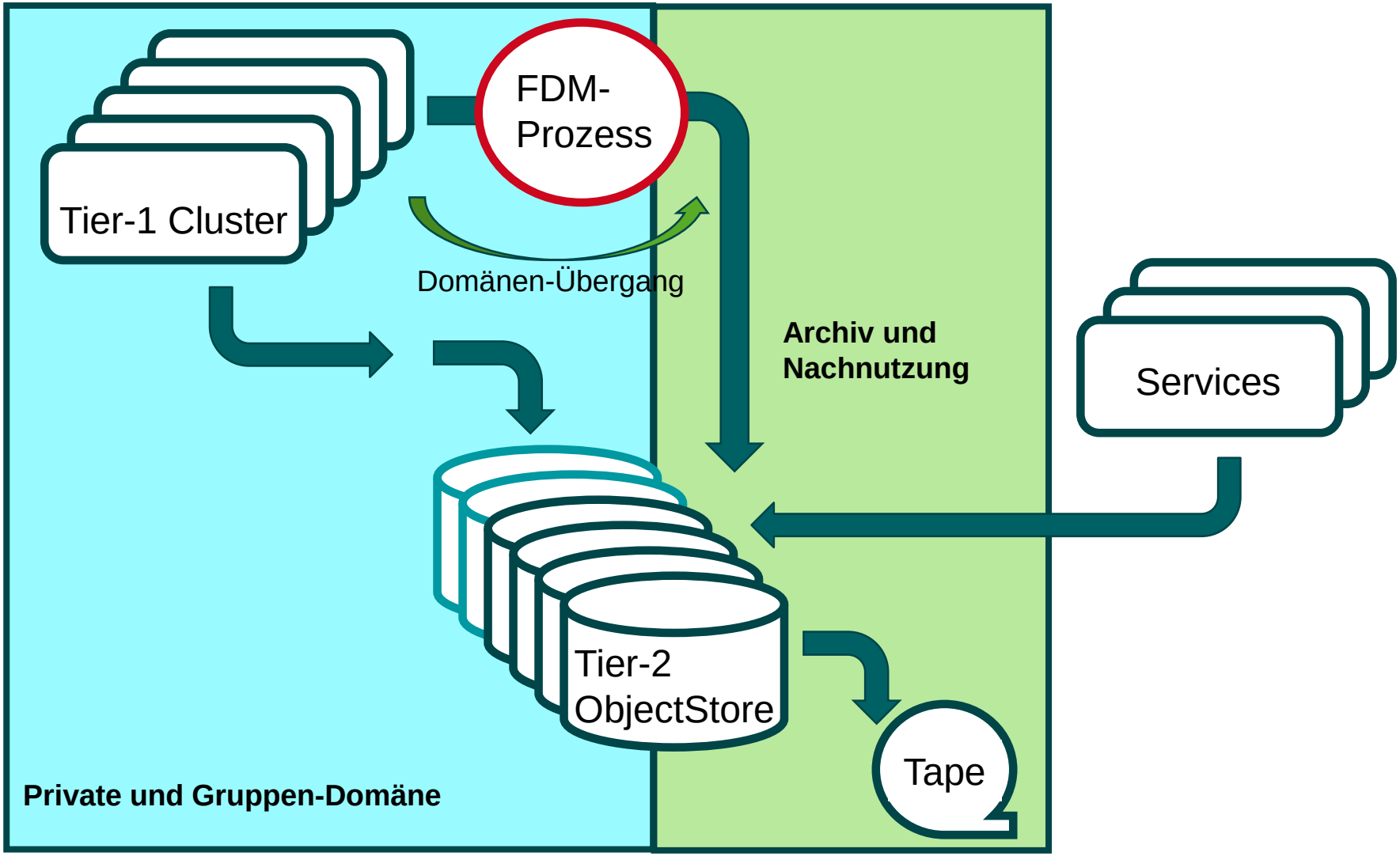
FD-Storage: 3-Tier-Speicherhierarchie (Archiv mit Speicherhierarchie)



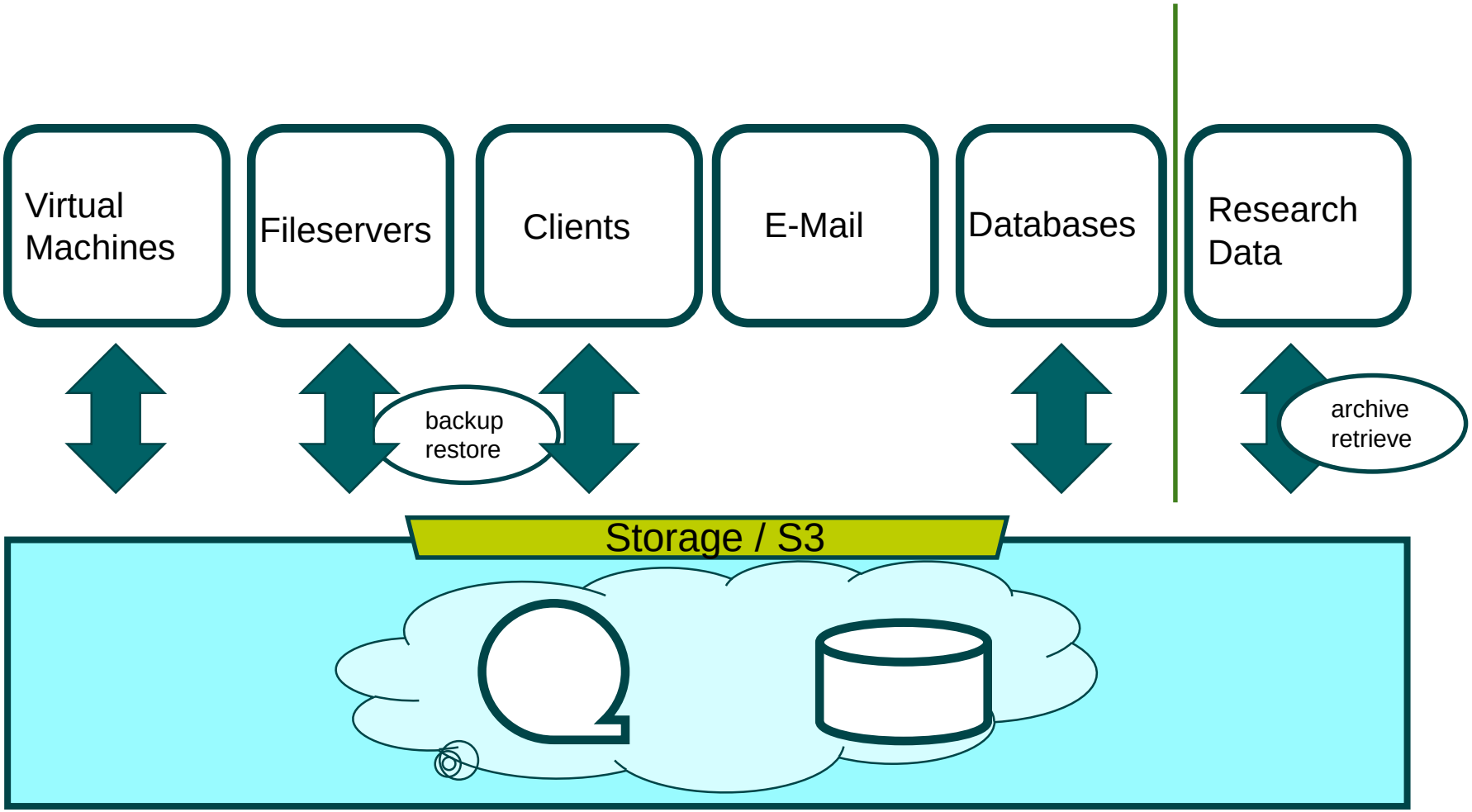
Folgerungen

- Objekt-Speicher wird zentrale Speicher-Instanz
- Sicherung gegen HW-Ausfälle auf Ebene des Objekt-Speichers
  - Sicherung gegen Ausfall / Fehler der Objekt-Speicher-Software nur für wenige selektierte Teilbereiche, z.B. durch Extern-Replikation
- Tape/kostenoptimierter Speicher als Extension / Kopier-Ziel des Objekt-Speichers
  ➔ Kontrolle bei Obj.

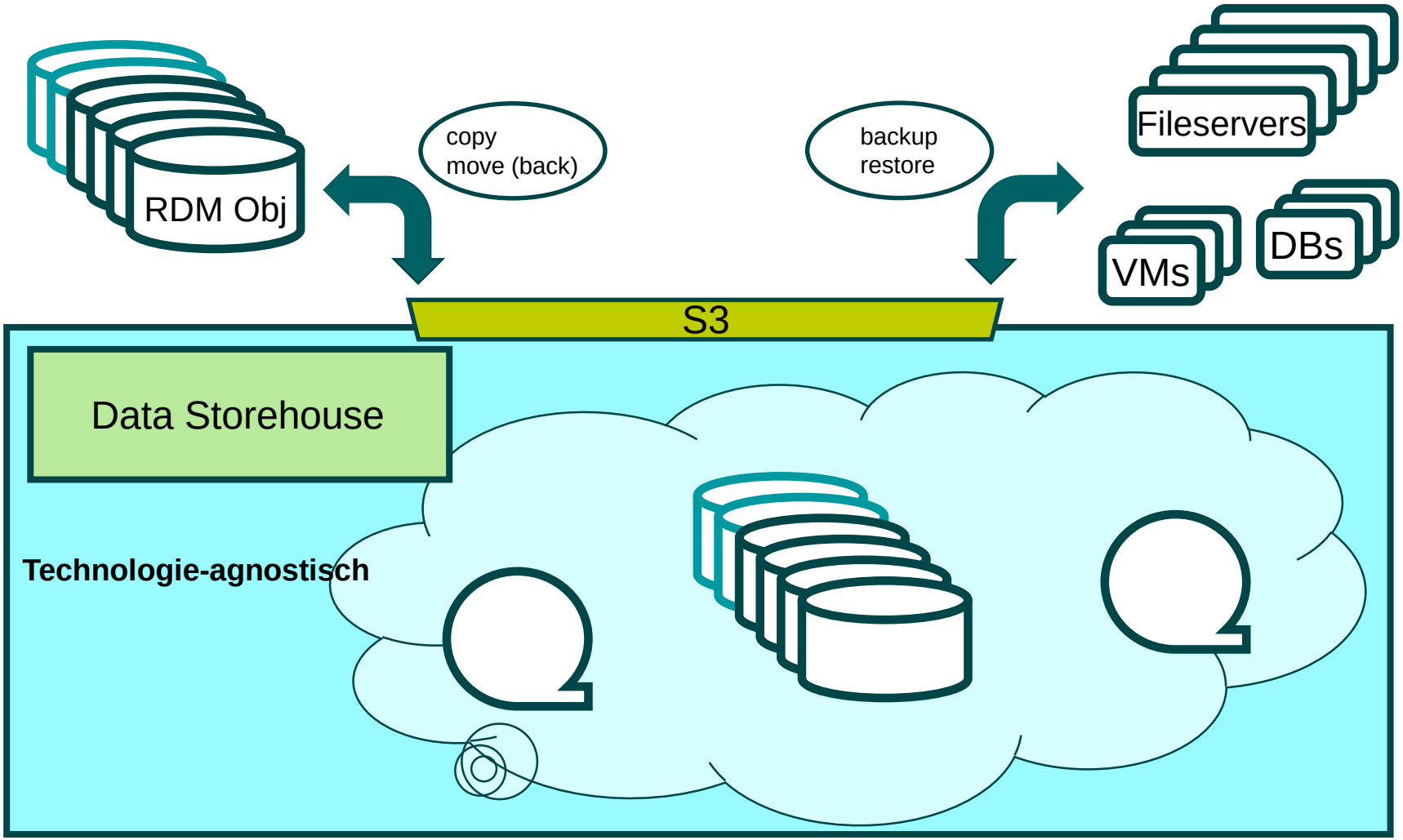
Object Store als zentrale Speicherinstanz



Data Storehouse: hochvolumiger sicherer Speicher



Data Storehouse: hochvolumiger sicherer Speicher



Folgerungen

- Sicherung von Endgeräten z. B. über (private) Cloud
- Sicherung von dezentralen Servern drastisch reduziert; Infragestellung der bisherigen Grundversorgung mit „Backup“

➔ Glaubwürdige Umsetzung von Backup als Transitionsstrategie (NRW TSM-Antrag 2014)

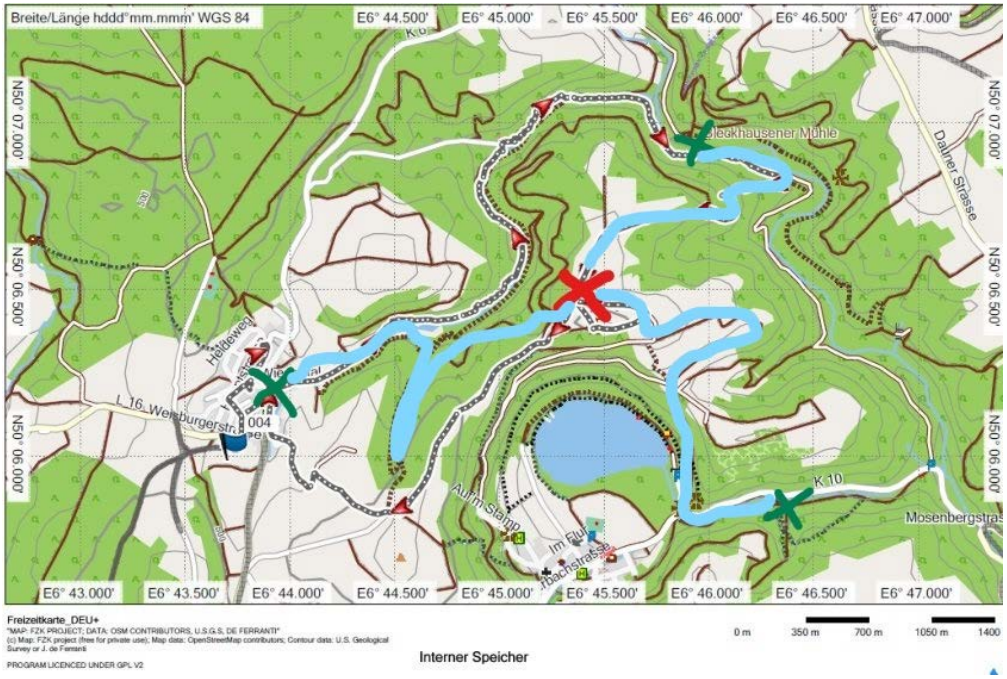
Zusammenfassung

- **Daten als Grundlage von Wissenschaft**
  ➔ Sichere Datenhaltung wird immer wichtiger
- **Objekt-Speicher: Abstraktion zwischen Anwendungen und Speicherhierarchie**
- **Differenzierter Einsatz von Speichermedien und –Methoden**
- **Klassisches Backup als (notwendige) Nischenlösung**

Sichere Datenhaltung im Wandel

IT Center | RWTH AACHEN UNIVERSITY

Vielen Dank für Ihre Aufmerksamkeit



IT Center | RWTH AACHEN UNIVERSITY