Datenmanagement in der Cloud für den Bereich Simulationen und Wissenschaftliches Rechnen

Peter Reimann, Tim Waizenegger, Matthias Wieland, Holger Schwarz

Institute für Parallele and Verteilte Systeme (IPVS)
Universität Stuttgart

2. Workshop Data Management in the Cloud auf der 44. GI-Jahrestagung
23. September 2014, Stuttgart, Deutschland

IPVS

Simulation des Knochenwachstums

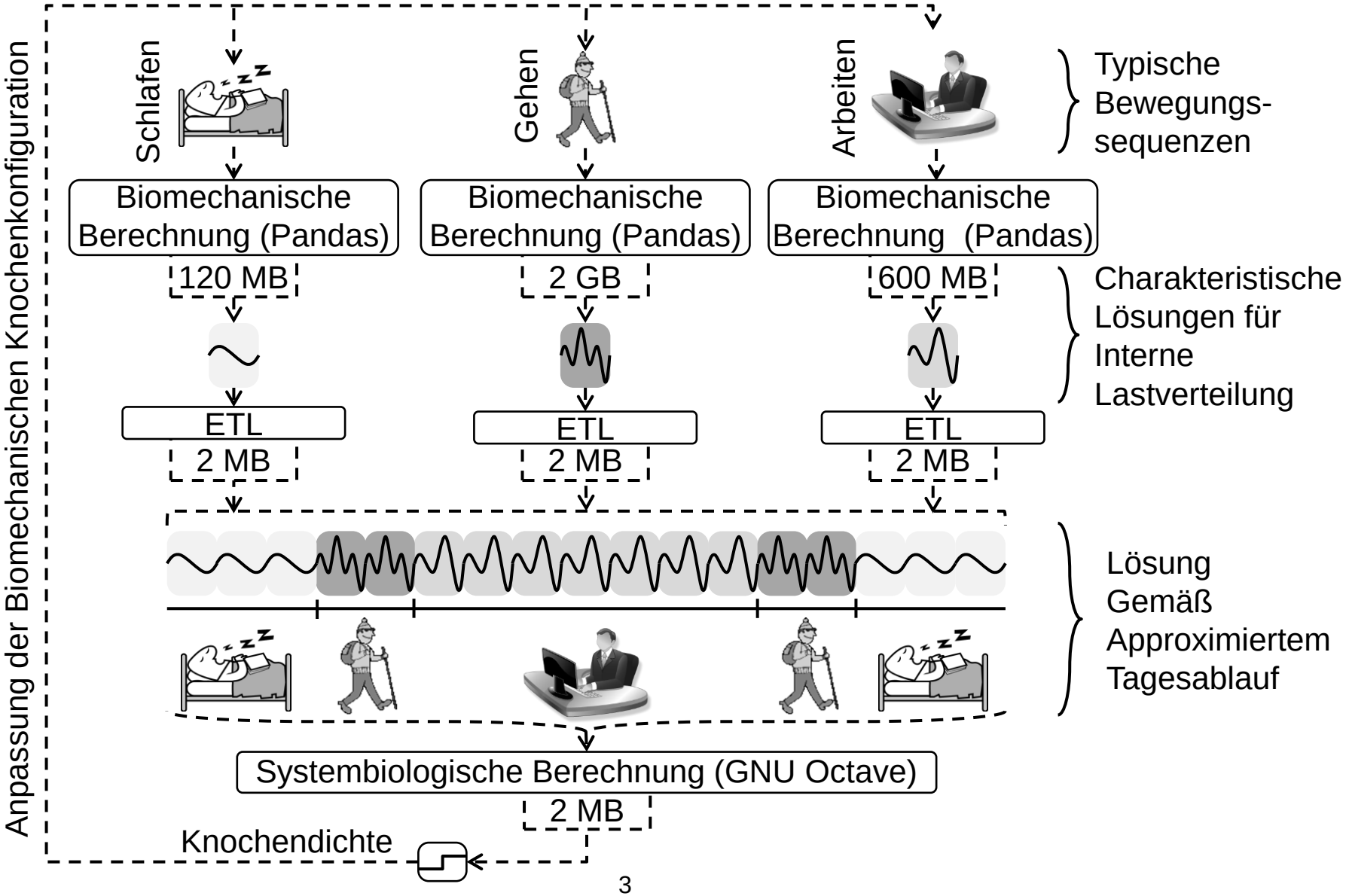
Biomechanische Berechnung (Pandas)

- Rechnet externe Last auf Knochen in interne Lastverteilung im Knochengewebe um

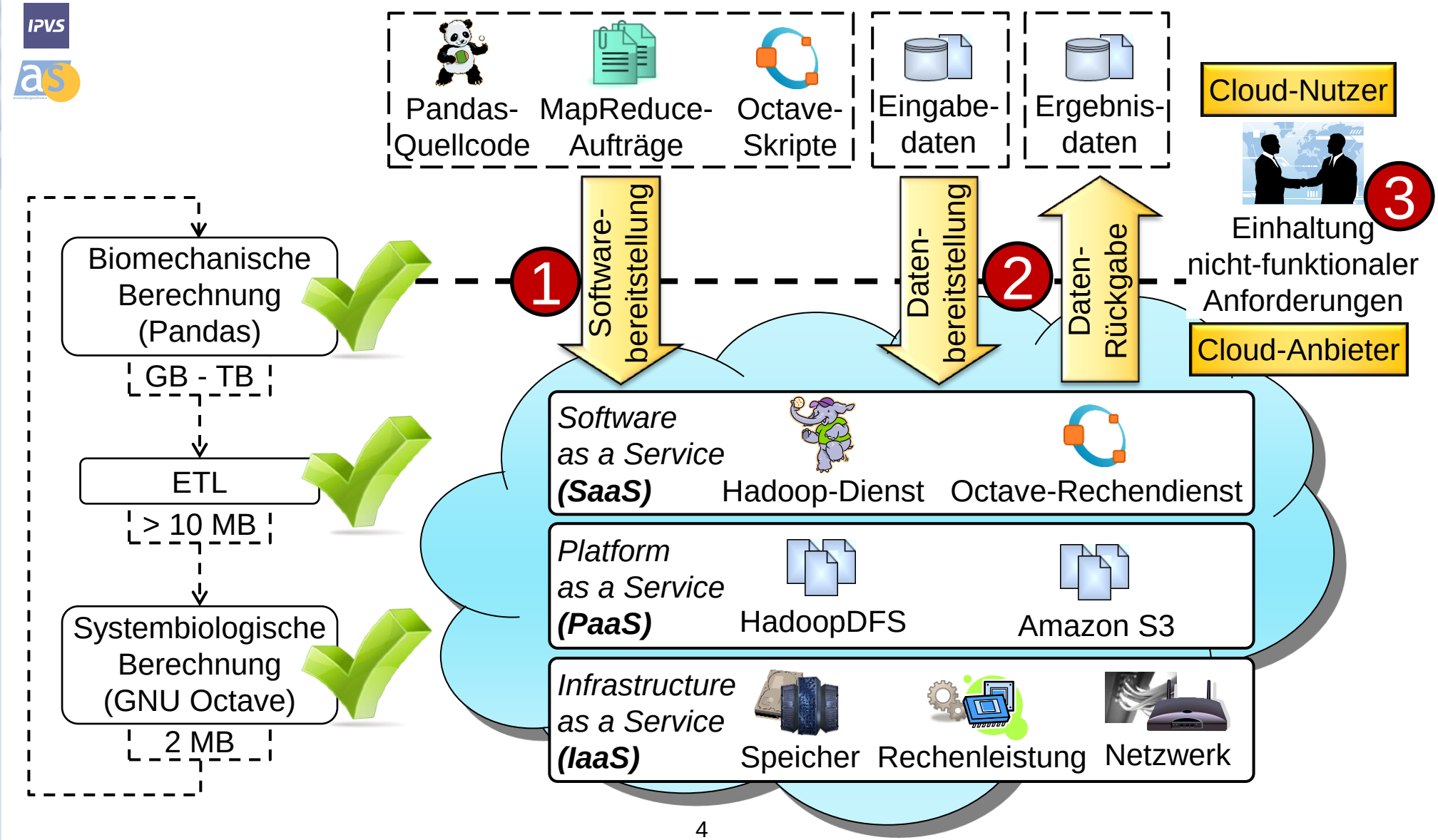
- Interne Lastverteilung beeinflusst Interaktion von Zellen im Gewebe
- Art der Zellinteraktion führt entweder zu Abbau oder zu Bildung von Knochengewebe

Systembiologische Berechnung (GNU Octave)

Simulation des Knochenwachstums



Umsetzung der Simulation in der Cloud



IPVS

as

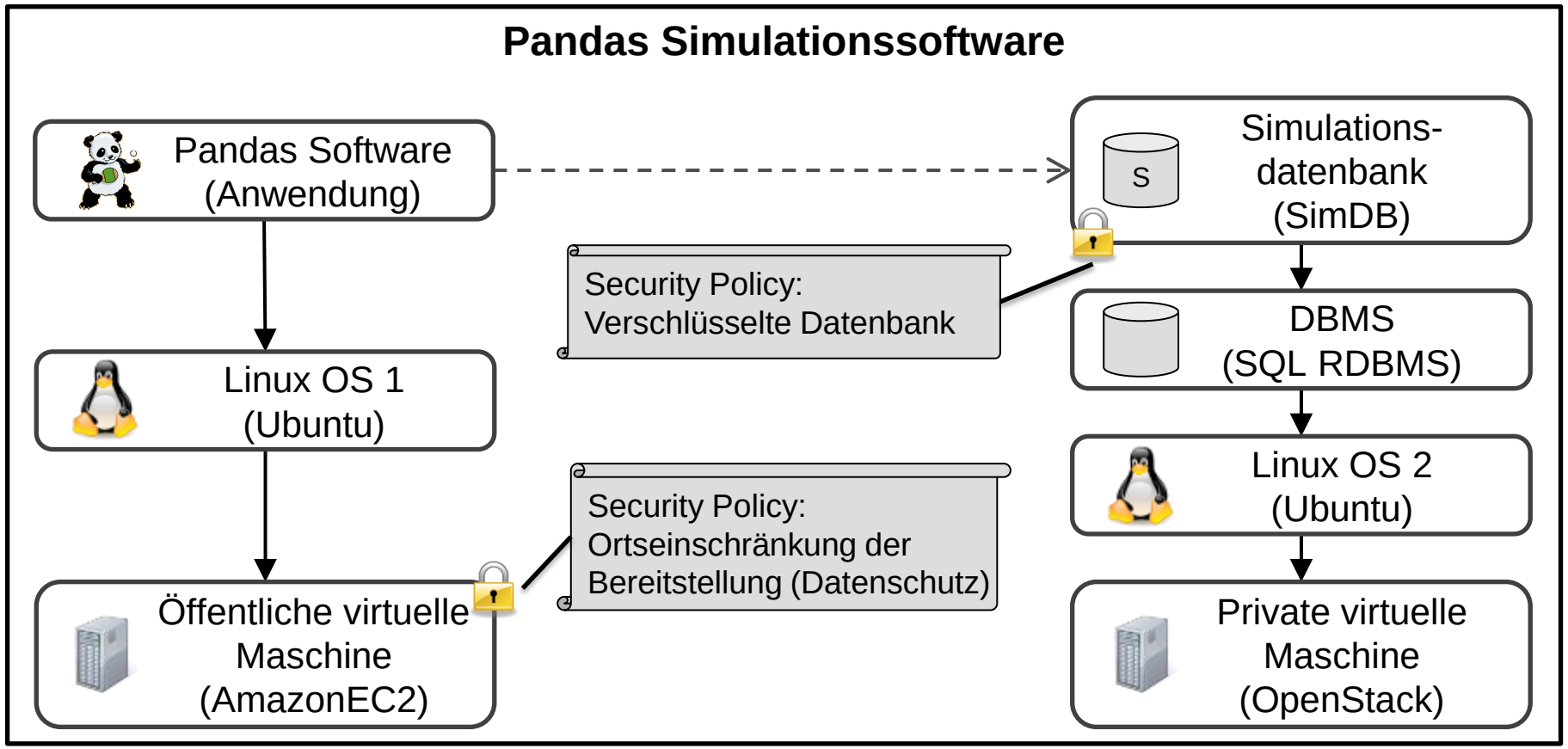
Agenda

- Motivationsszenario Knochensimulation
- Softwarebereitstellung
  - Einsatz des OASIS Standards TOSCA für Pandas-Software

    T*opology and* ***O****rchestration* ***S****pecification for* ***C****loud* ***A****pplications*
- Datenbereitstellung und Datenrückgabe
  - Diskussion verschiedener Alternativen
  - Einsatz des SIMPL-Rahmenwerks für Pandas

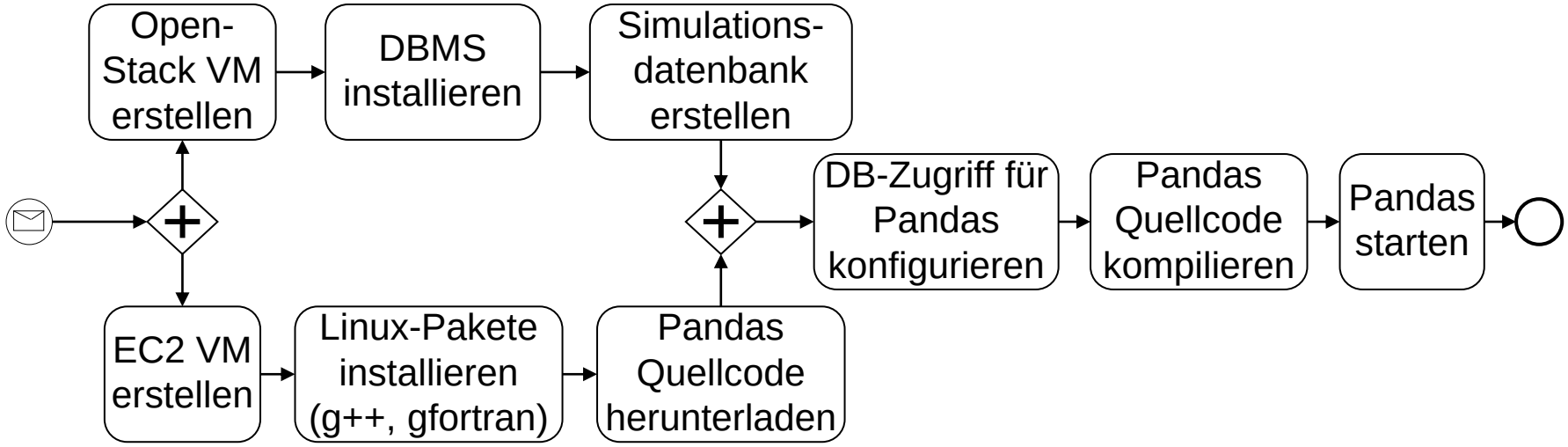
    S*imTech –* ***I****nformation* ***M****anagement,* ***P****rocesses, and* ***L****anguages*
- Einhaltung nicht-funktionaler Anforderungen
  - Diskussion von Möglichkeiten zur Definition und Einhaltung relevanter Anforderungen

TOSCA Diensttopologie für Pandas



TOSCA Pläne für Bereitstellung und Verwaltung einer Cloud-Anwendung

- Deployment-Plan für automatisierte Bereitstellung der Pandas-Software



- Management-Pläne für deren Verwaltung zur Laufzeit
  - In unserem Kontext insbesondere für Datenbereitstellung und Datenrückgabe (siehe nachfolgende Folien)

IPVS

Datenbereitstellung: Anforderungen

- Generische Lösung zur Kopplung beliebiger Simulationen
  - Unterstützung einer Vielzahl an heterogener Datenformaten und Datenmanagementoperationen
- Abstraktionsunterstützung für Wissenschaftler
  - Wissenschaftler haben keine hohe Expertise im Bereich des Datenmanagements
  - Sollten nicht mit komplexen Implementierungsdetails überfordert werden
- Integration der Datenbereitstellung in TOSCA wünschenswert
  - Gleiche Lösung für Software- und Datenbereitstellung

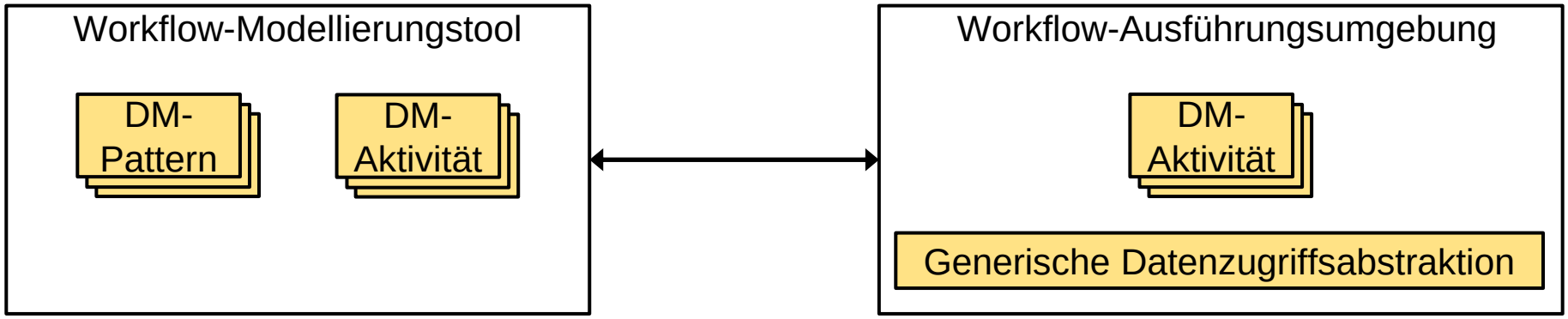
Alternativen für die Datenbereitstellung

- ETL-Tools (Extraktion, Transformation, Laden)
  - ✓ Bieten vielfältige und damit generische Möglichkeiten zur Umsetzung von Datenmanagementoperation
  - – Keine für Wissenschaftler adäquate Abstraktionsunterstützung
  - – Integration in TOSCA schwierig
- Erweiterungen von Workflow-Sprachen zur Einbettung von Datenbankanweisungen (z.B. BPEL/SQL)
  - ✓ Integration in TOSCA-Pläne über Workflows leicht möglich
  - – Meist eingeschränkt auf bestimmte Datenressourcen und damit nicht generisch
  - – Ebenso keine adäquate Abstraktionsunterstützung

IPVS

AS

SIMPL-Rahmenwerk



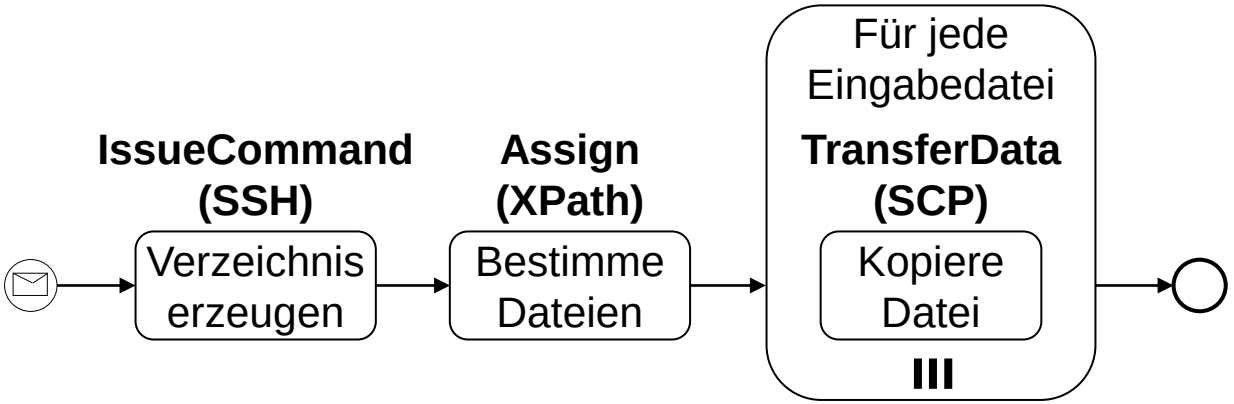
- Generischer Datenzugriff und generische Datenmanagement-Aktivitäten
  - ✓ Keine Einschränkung auf bestimmte Datenressourcen / Datenformate
  - ETL-Tools bieten häufig noch mehr (komplexe) Operationen als die Datenressourcen
- ✓ Für Wissenschaftler geeignete Abstraktionsunterstützung über typische Datenmanagement-Patterns
- ✓ Integration in TOSCA-Pläne über Workflows leicht möglich

➔ Offener Punkt: Wie kann SIMPL mit ETL-Tools kombiniert werden?

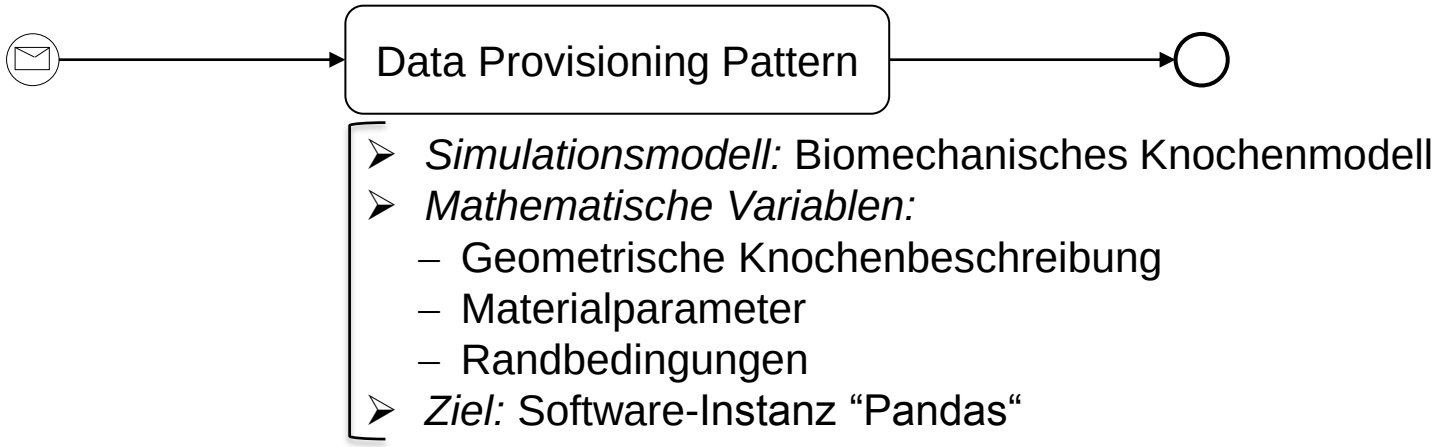
IPVS

SIMPL-Datenmanagement-Aktivitäten in TOSCA Management-Plänen

- TOSCA-Plan zur **Datenbereitstellung** für Pandas



Patterns als Abstraktionsunterstützung



- Reduktion der sichtbaren Workflow-Aktivitäten von 4 auf 1
- Modellierung dieser Aktivität deutlich weniger komplex
  - Hauptsächlich Begrifflichkeiten aus Simulationsmodellen, mit denen Wissenschaftler bereits vertraut sind

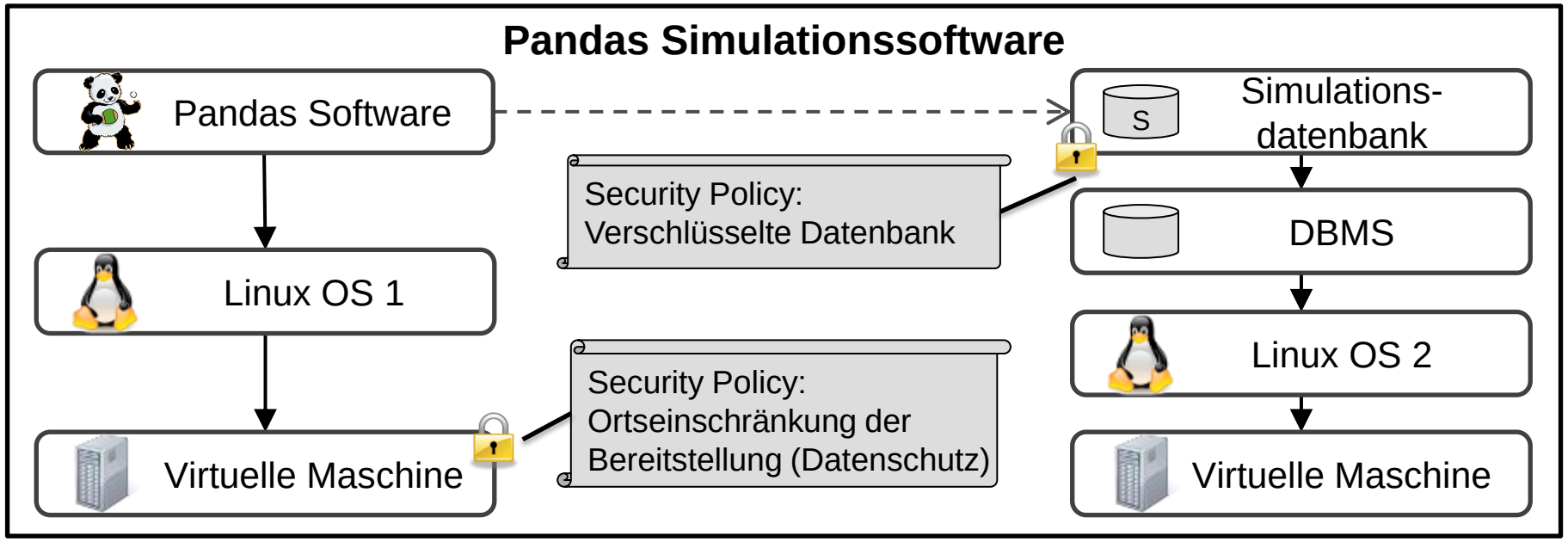
IPVS

as

Für Simulationen Relevante Nicht-Funktionale Anforderungen

- Datensicherheit
  - Schutzwürdiges geistiges Eigentum oder gesetzliche Bestimmungen
- Datenqualität
  - Mathematische Simulationsmodelle sowie deren numerische Implementierungen bringen Ungenauigkeiten mit ein
- Effizienz und Optimierungsmöglichkeiten
  - Stichpunkt Big Data
- Reproduzierbarkeit und Nachvollziehbarkeit
  - Simulation erneut durchführbar machen
  - Logging / Monitoring / Provenance

Datensicherheit: Policy4TOSCA-Rahmenwerk



- Policies ermöglichen frei definierbare Anforderungen
- Einhaltung von Datensicherheitsanforderungen über
  - Auswahl passender Anwendungskomponenten
  - Anpassung der TOSCA Deployment- bzw. Management-Pläne

Effizienz und Optimierungsmöglichkeiten

- Untersuchung verschiedener Optimierungsmöglichkeiten in Bezug auf Tauglichkeit für Simulationen
  - Änderungen von TOSCA Diensttopologien
    - Z.B. Datenbank-Stack für Pandas in öffentlicher Cloud-Umgebung
    - Dann aber negative Auswirkungen auf Datensicherheit möglich

    ➔ Methoden zur flexiblen Reaktion auf sich konkurrierende Anforderungen
  - Ansätze zur Workflow-Restrukturierung
    - Z.B. Optimierungen von Workflows mit eingebetteten SQL-Anweisungen

      Dissertation von Marko Vrhovnik, Universität Stuttgart, 2011

    ➔ Für generische Lösung Erweiterung auf andere Datenressource
  - MapReduce / Big Data
    - Z.B. für komplexe ETL-Operationen als Hadoop-Dienst in der Cloud

IPVS

Zusammenfassung

- Cloud-Infrastrukturen zur Umsetzung von Simulationen
  - Geeignet für große und variierende Datenvolumina sowie Rechenaufwände, speziell bei gekoppelten Simulation
- OASIS-Standard TOSCA zur Umsetzung der Softwarebereitstellung für Simulationen in der Cloud
- Alternativen für die Datenbereitstellung und –rückgabe
  - Integration des SIMPL-Rahmenwerks mit TOSCA
  - Offener Punkt: Kombination von SIMPL mit ETL-Tools
- Definition und Einhaltung nicht-funktionaler Anforderungen
  - Policy4TOSCA-Rahmenwerk
    - Einhaltung der Anforderungen hauptsächlich für Datensicherheit
  - Offener Punkt: Untersuchung und Integration weitere Anforderungen

IPVS

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT