Benjamin Vogler
Hochschule RheinMain
Master Informatik
24. Juli 2014

# Halbautomatisierte und modellgestützte Leistungsanalyse von Speichersystemen

Master Kolloquium

## Agenda

- Motivation
- Grundlagen
- Ansatz
- SVC Speichersystem
- Modellierung
- Automatisierung
- Testergebnisse
- Zusammenfassung

- Labor für Verteilte Systeme Hochschule RheinMain
- Industriepartner System Vertrieb Alexander
  - IBM SAN Volume Controller (SVC)
  - Internes Prognose-Tool zur Leistungsvorhersage (DiskMagic)
  - Management-Umgebung BVQ
- Projekt OntoStorM
  - Ontologie-basiertes Speicher-Management
- Gefördert über Mittel aus Hessen LOEWE 3
- Bearbeitungszeit 6 Monate

Problemstellung

- Steigende Komplexität der Speichersysteme
- Vorhersage der Leistungsfähigkeit eines Speichersystems
- Einhaltung von SLAs zwischen Betreiber und Kunden

Ziele

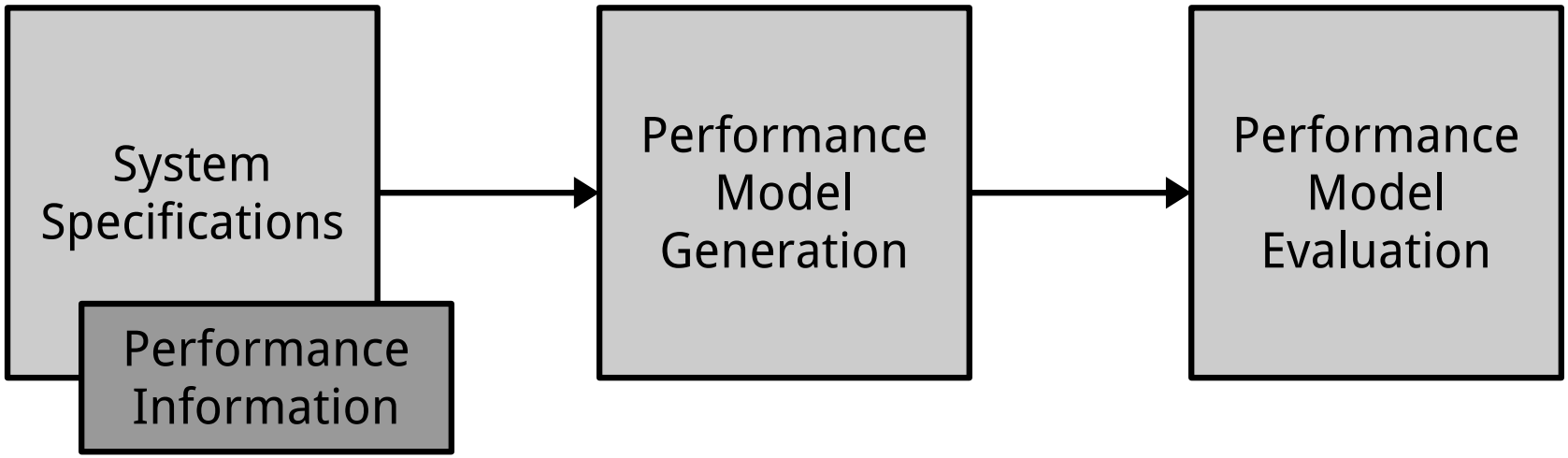
- Modellgestützte Leistungsanalyse SVC-basierter Speichersysteme
- Vorhersage von Leistungsindizes für angenommene Szenarien
- Domänenspezifische Modelle
- Integration in die Management-Umgebung BVQ

Lösungsansatz

- Adaption eines Software Performance Engineering Ansatzes

## Software Performance Engineering [Smi93]

- Erstellung eines Systemmodells
- Erweiterung um Leistungsinformationen
- Transformation in ein Leistungsmodell
- Auswertung des Leistungsmodells



UML

- System-Modellierungssprache
- OMG Standard 2.4.1
- Grafische und maschinenlesbare (XMI) Repräsentation
- Eclipse Papyrus Modellierungsumgebung

MARTE

- UML Profil für Echtzeitsysteme und eingebettete Systeme
- OMG Standard 1.1
- Anpassung an Speicher- und Leistungsdomäne über Stereotypen (HRM/PAM/GQAM)
- Variabilität über Sprachausdrücke (VSL)

## Sprachen zur Modelltransformation und Leistungsmodellierung

QVT (Query/View/Transformation)

- Programmatische Transformationssprache
- OMG Standard 1.1
- Ausprägung imperativ (QVT-o)/deklarativ (QVT-r)
- Aktuelle Implementierung nur für QVT-o vorhanden (Eclipse)

LQN (Layered Queueing Model)

- Leistungs-Modellierungsprache
- Geschichtete Warteschlangennetzwerke (Client-Server-Prinzip)
- Entwickelt an der Universität Carleton, Ottawa [Fra12]
- Grafische und maschinenlesbare (XML) Repräsentation
- Proprietäre Solver verfügbar

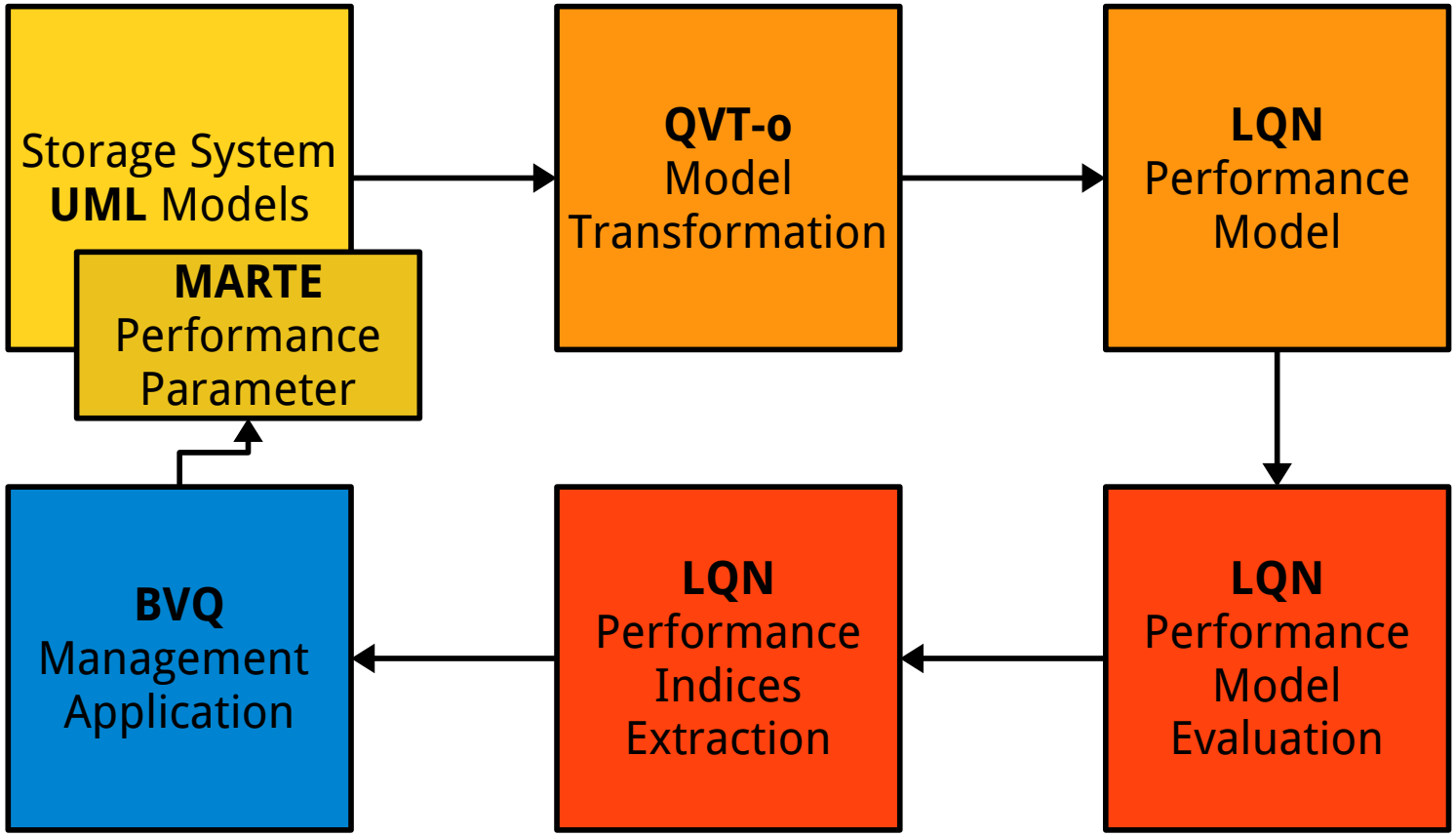
Machbarkeitsstudie [Vog13]

- Einfache Systemmodelle in UML und MARTE
- Transformation mit QVT-o
- Evaluation über LQN Solver
- Mangelnder Detailgrad
- Unzulängliche Testergebnisse
- Keine Automatisierung

## Methoden aus der Literatur

- Aizikowitz u.a.: *Component-Based Performance Modeling of a Storage Area Network* [Aiz+05]
  - Programmatisches Systemmodell
  - Evaluation über Simulationsumgebung
  - Hoher Detailgrad des SVC Systemmodells
  - Mangelnde Nachvollziehbarkeit der Modelle
- Tribastone u.a.: *Performance Prediction of Service-Oriented Systems with Layered Queueing Networks* [TMW10]
  - Systemmodelle in UML und MARTE
  - Evaluation über LQN
  - Verständliche Modelle
  - Blackbox-Modell Speichersystem
  - Manuelle Transformation

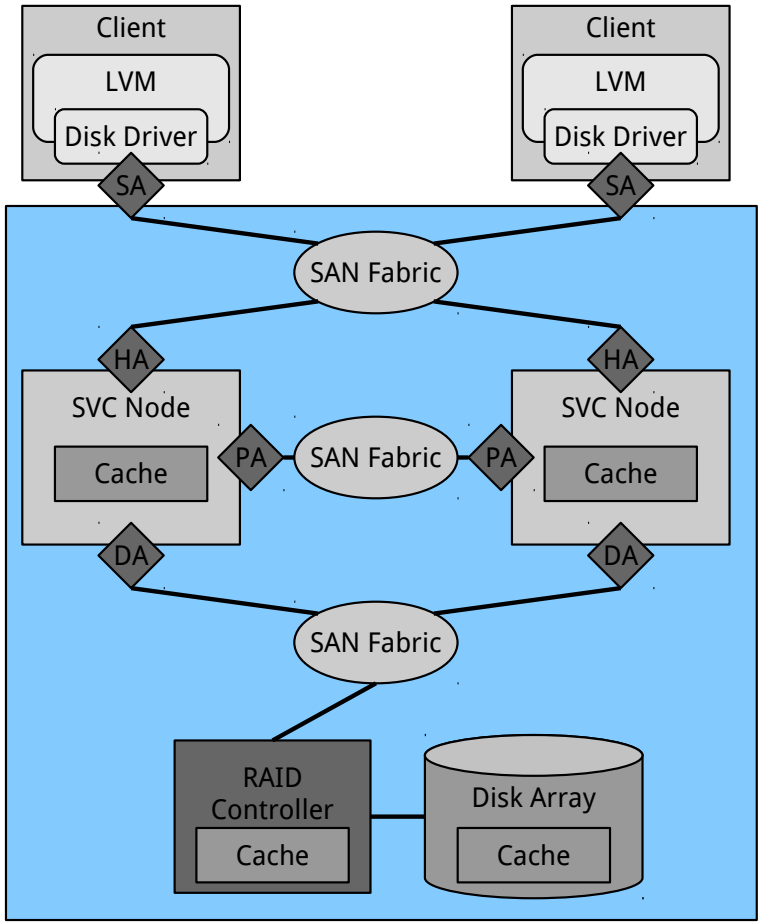
## Ausgewählte Methode aus dem Bereich Software Performance Engineering



## Struktur [Tat+10]

SVC Systemkomponenten

- Adapter (Storage/Host/Disk)
- SAN Fabric (Host/Peer/Disk)
- SVC Node
- Cache
- RAID Controller
- Laufwerke



Caching-Verhalten

- Read-Ahead
- Write-Back

RAID-Zugriffsverhalten [Why12]

- Abbildung Clientanfragen auf Disk Arrays
- Striping
- Parity Disk / Mirroring

HDD-Zugriffsverhalten [RW94]

- Zufällige bzw. sequentielle Lese- und Schreibzugriffe
- Seek Time / Rotational Latency / Head Switch Time

## Leistung, Last und Konfiguration

Leistungsindizes

- Antwortzeit [ms]
- Auslastung [prozentual]

Lastparameter

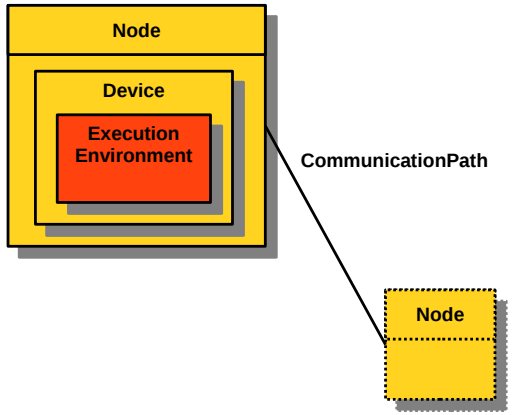
- Ankunftsrate [IOPS]
- I/O Verhalten (Lesen/Schreiben) [prozentual]
- I/O Verhalten (Sequentiell/Zufällig) [prozentual]
- Cache Hit Ratio [prozentual]

Konfigurationparameter

- Dauer der Befehlsverarbeitung [ms]
- Bandbreite (Lesen/Schreiben) [MB/s]
- Abarbeitungszeit [ms]
- Multiplizität [integer]

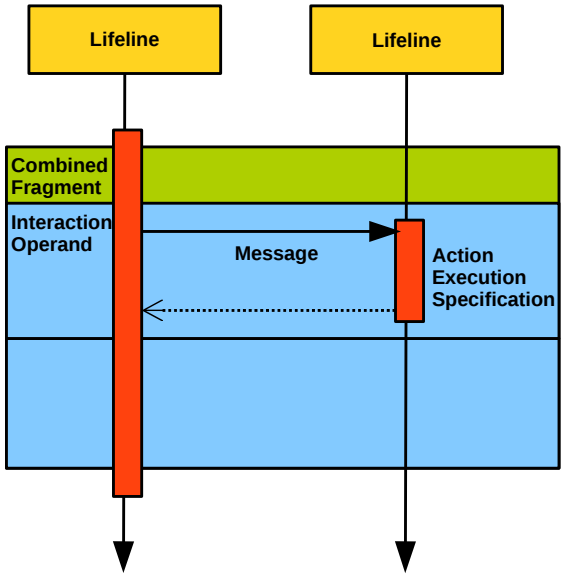
Struktur ⇒ Deployment-Diagramm

- Node
- Device
- Execution Environment




Verhalten ⇒ Sequenz-Diagramm

- Lifeline
- Action Execution Specification
- Combined Fragment
- Interaction Operand
- Message (sync/async)



Anwendungsfall ⇒ Use-Case-Diagramm

- Actor
- Usecase
- VSL Definitionen

Last ⇒ MARTE

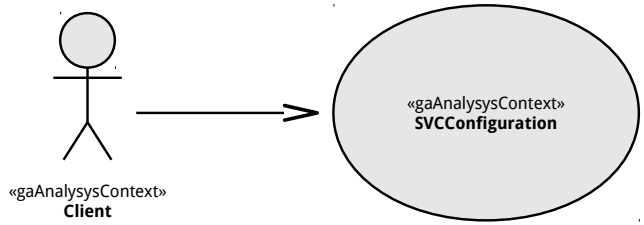
- PAM / GQAM Stereotypen

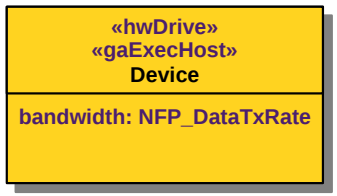
Leistung ⇒ MARTE

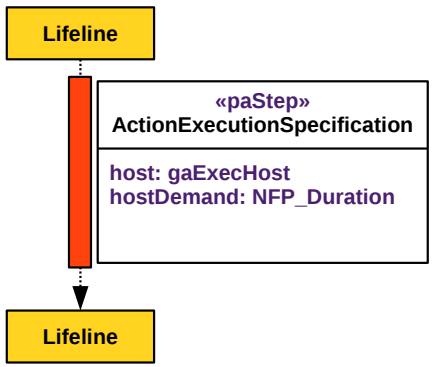
- PAM / GQAM Stereotypen

Konfiguration ⇒ MARTE

- HRM / GQAM Stereotypen



Abarbeitung

- Processor
- Task

Dienste

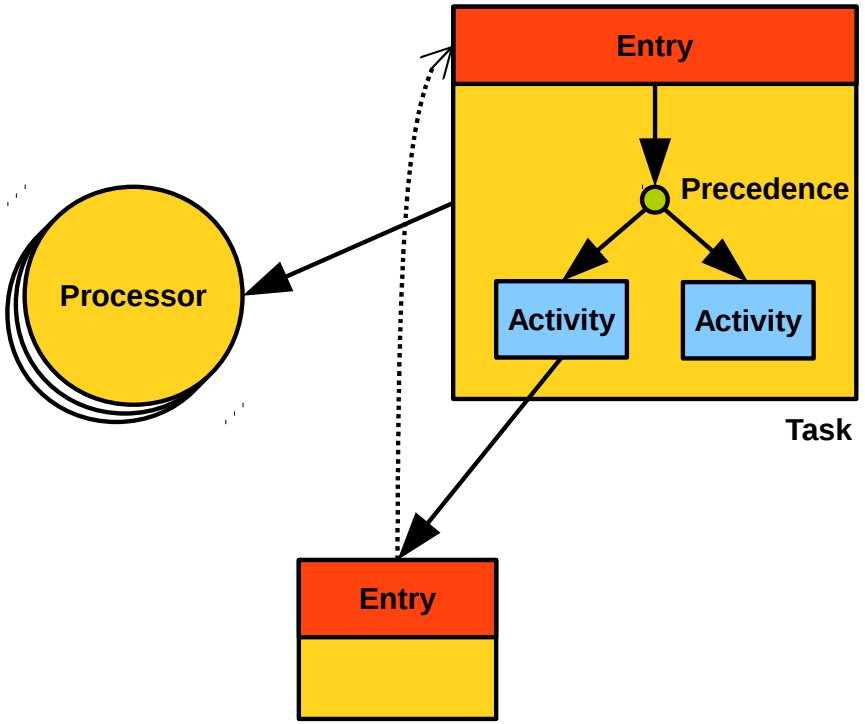
- Entry (Phase/Graph)

Ablauf

- Activity
- Precedence
- Call (sync/async)

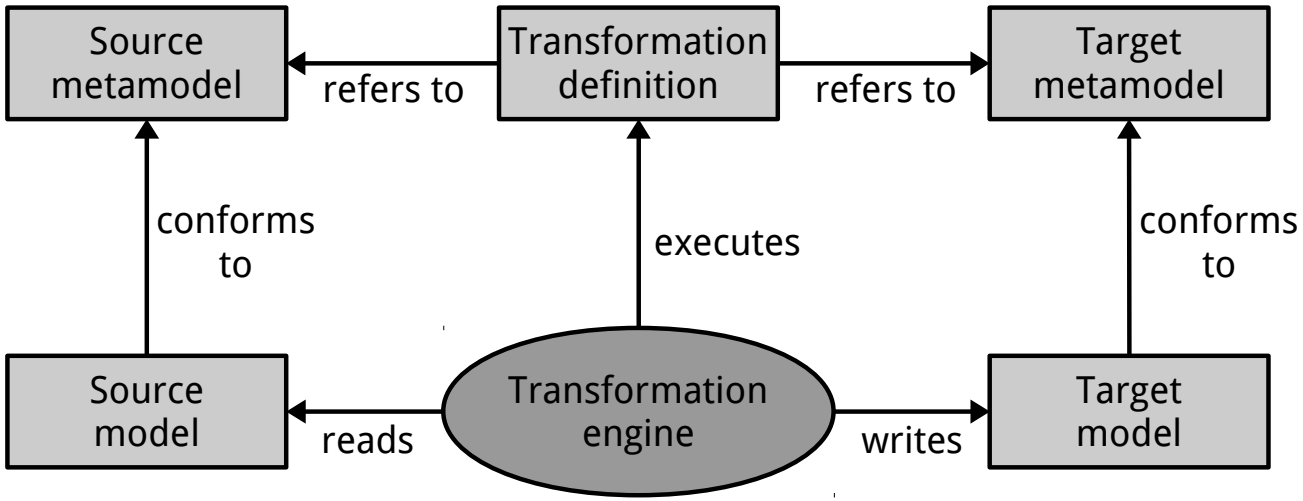
Last

- Reference Task



Transformation

- Import der Metamodelle
- Regeln: Metamodell-Ebene
- Ausführung: Modellinstanz-Ebene



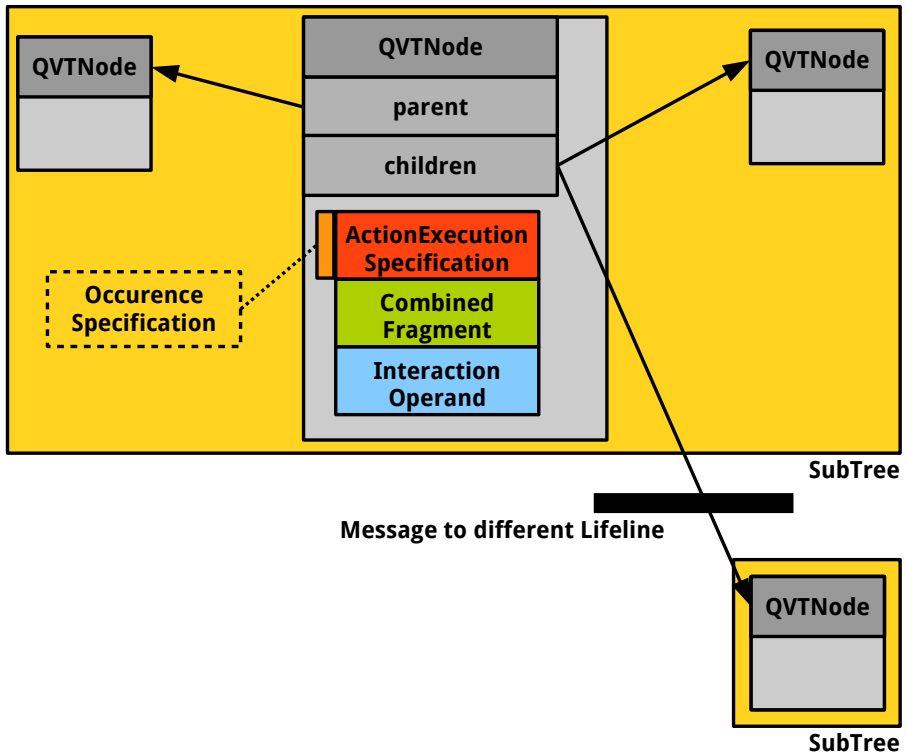
| **Systemmodell Element** | **Leistungsmodell Element** |
|---|---|
| Device<br>Lifeline | Task und Processor |
| ExecutionEnvironment<br>ActionExecutionSpecification | Entry oder Activity |
| CombinedFragment | Precedence |
| InteractionOperand | Activity |
| Message | Call, reply-Entry |

Herausforderung

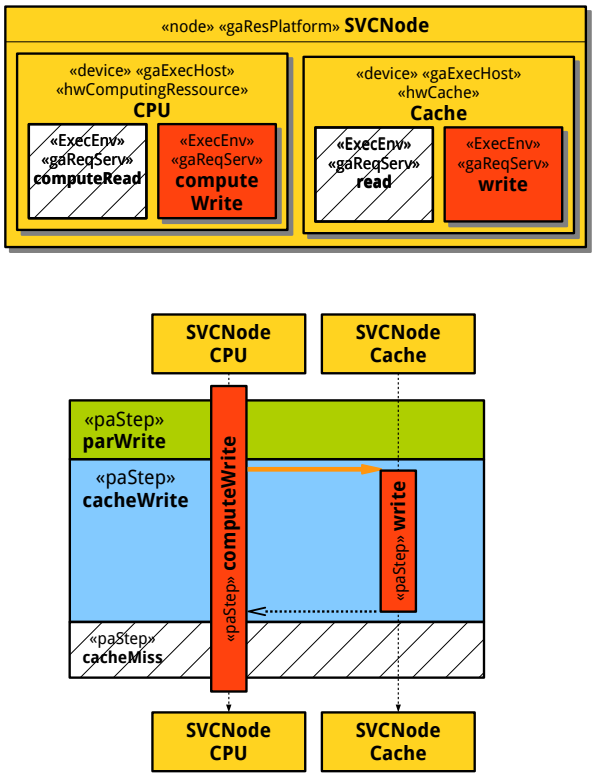
- 3 Quellmodelle $\Rightarrow$ 1 Zielmodell
- Unterschiedliche Modellstruktur
- Umwandlung Parameter

Zwischenmodell

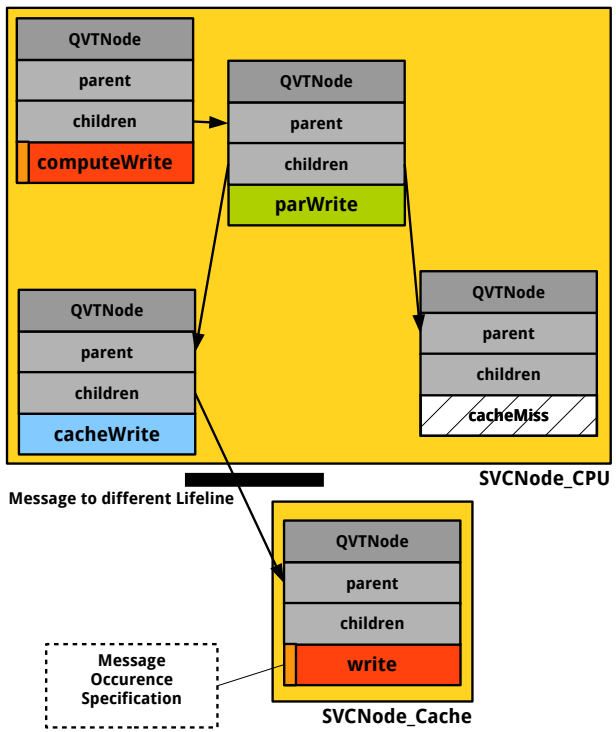
- Baumdiagramm
- Teilbäume pro Systemkomponente
- Knoten als Wrapper für Dienste/Ablauf



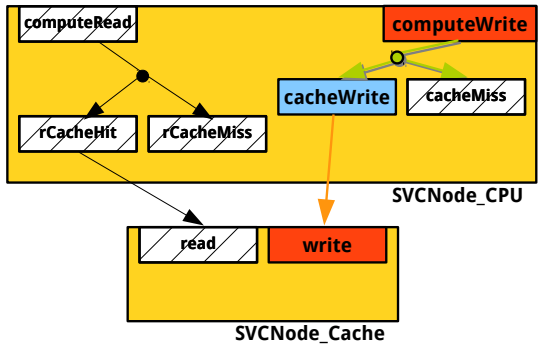
# Beispiel Schreiben SVC Node/Cache




Systemmodell



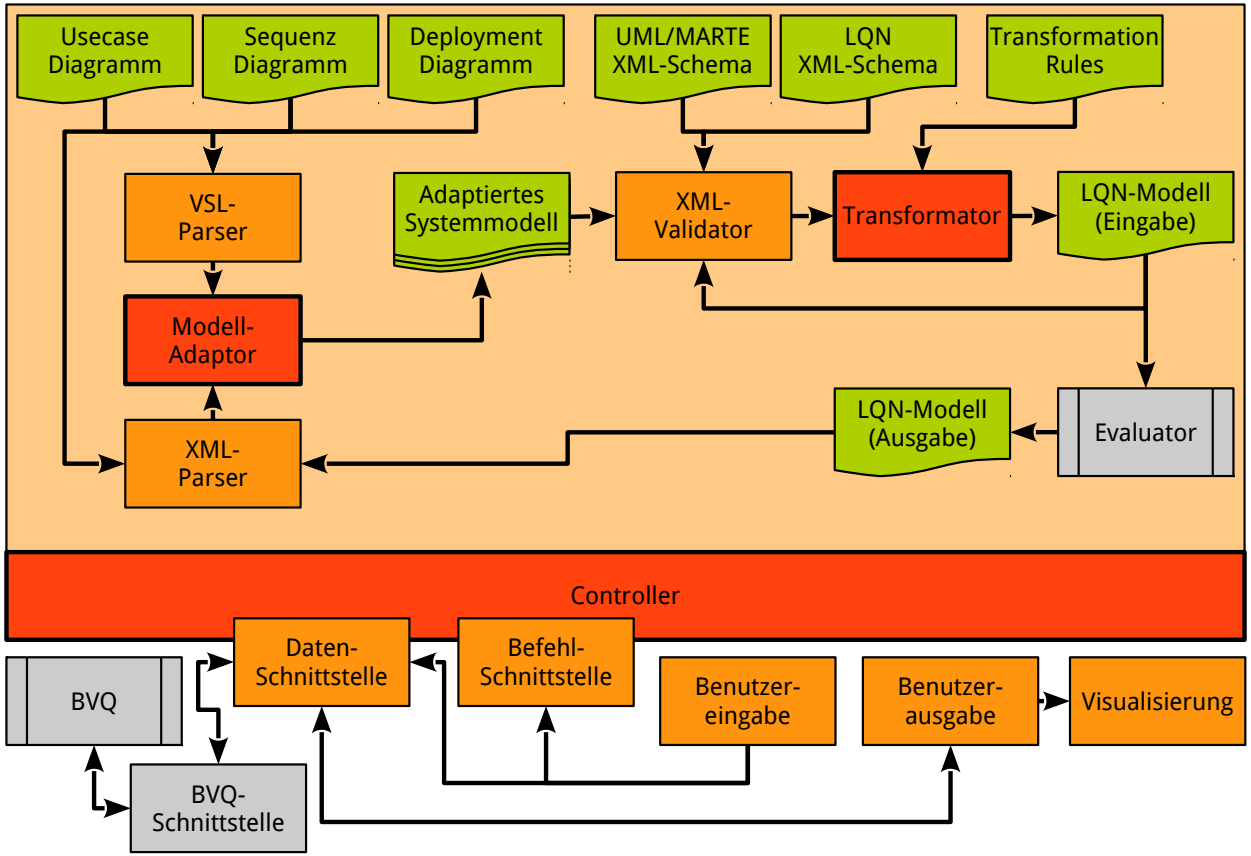

Zwischenmodell



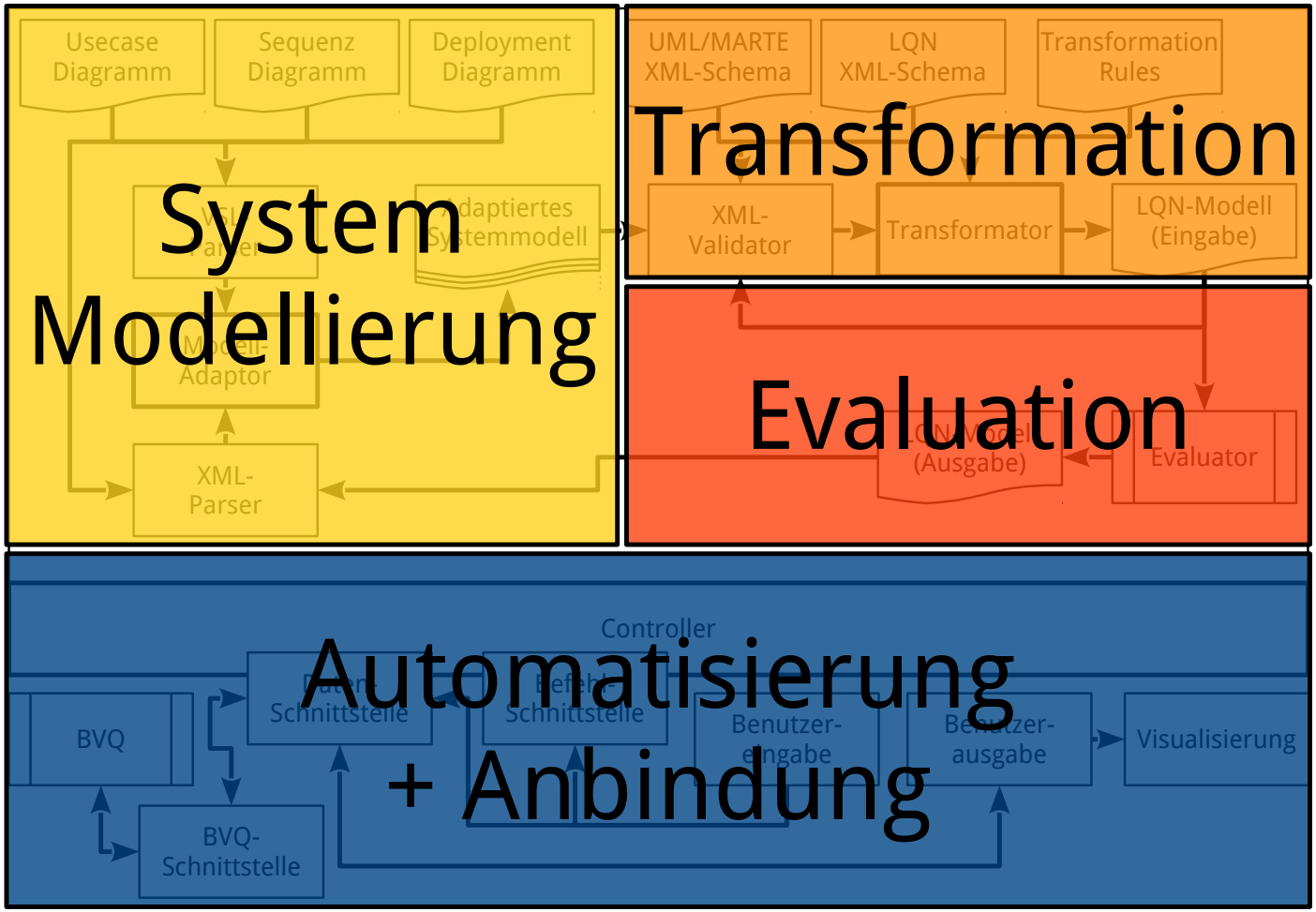

Leistungsmodell

## Anwendungskonzept Grobarchitektur

- Verbindung der Werkzeuge
- Automatisierung des SPE-Prozesses

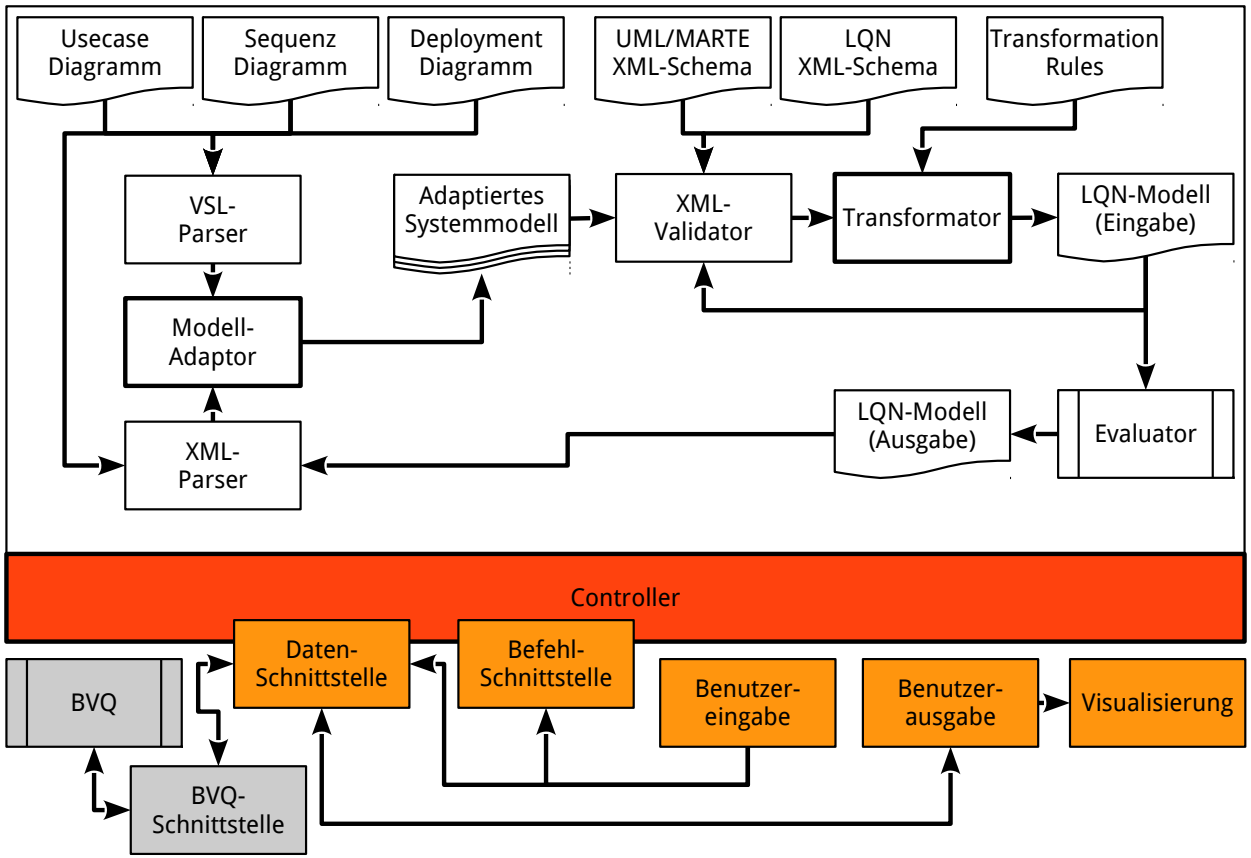


## Anwendungskonzept SPE-Prozess



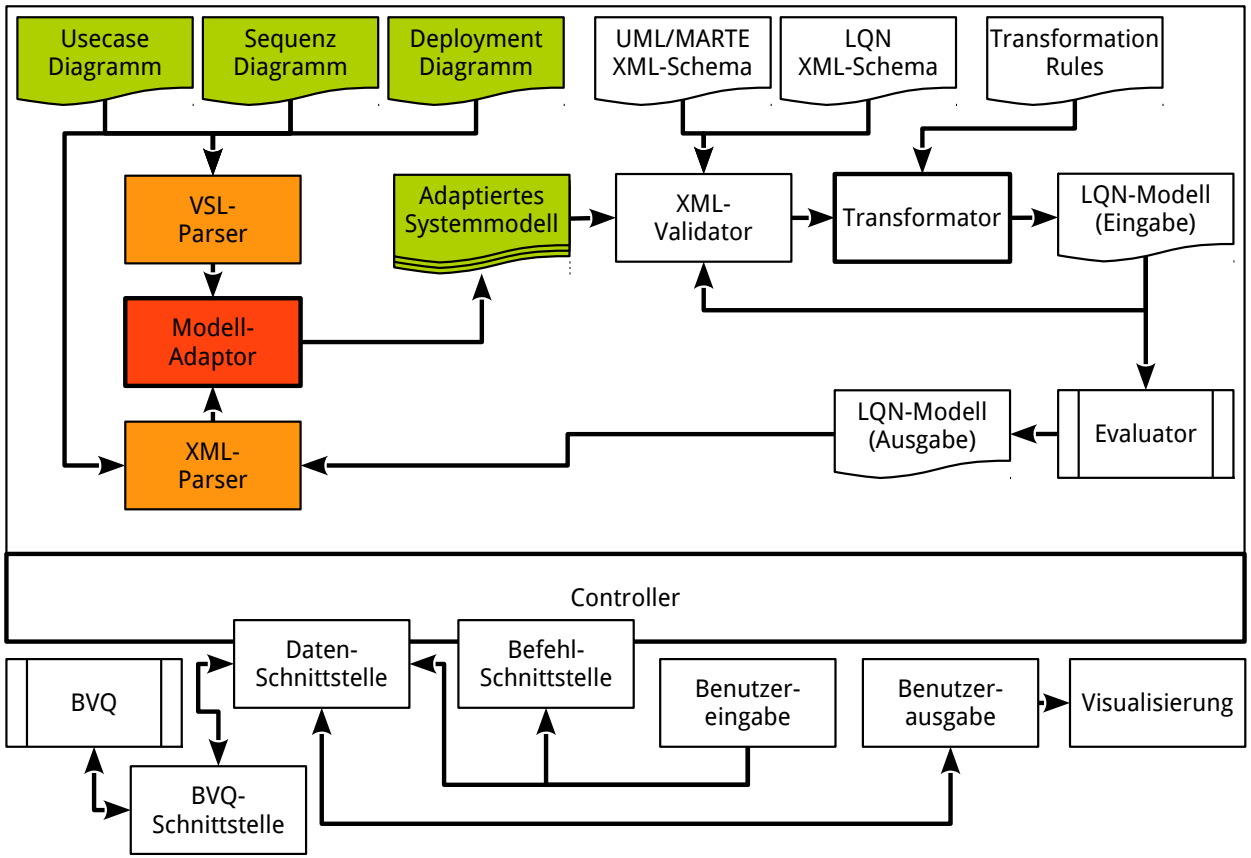
## Controller

- Datenschnittstelle zu BVQ oder Benutzereingaben
- Befehlschnittstelle zur Ausführung
- Benutzerausgabe oder Rückführung in BVQ



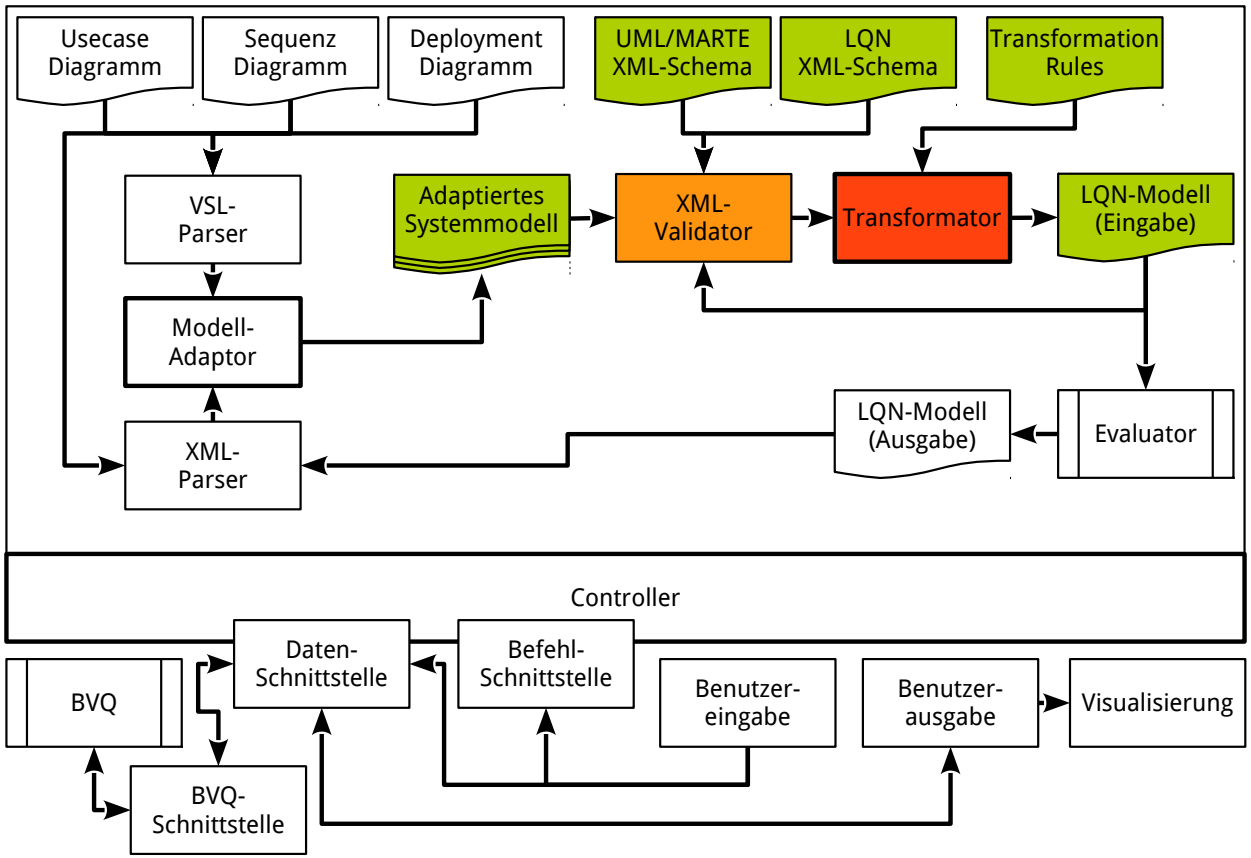
## Modell-Adaptor

- Adaption der Grundmodelle an angenommenes Szenario
- XML Parser zum Einlesen des UML-Modells
- VSL Parser zum Einlesen der Last- und Konfigurationsparameter



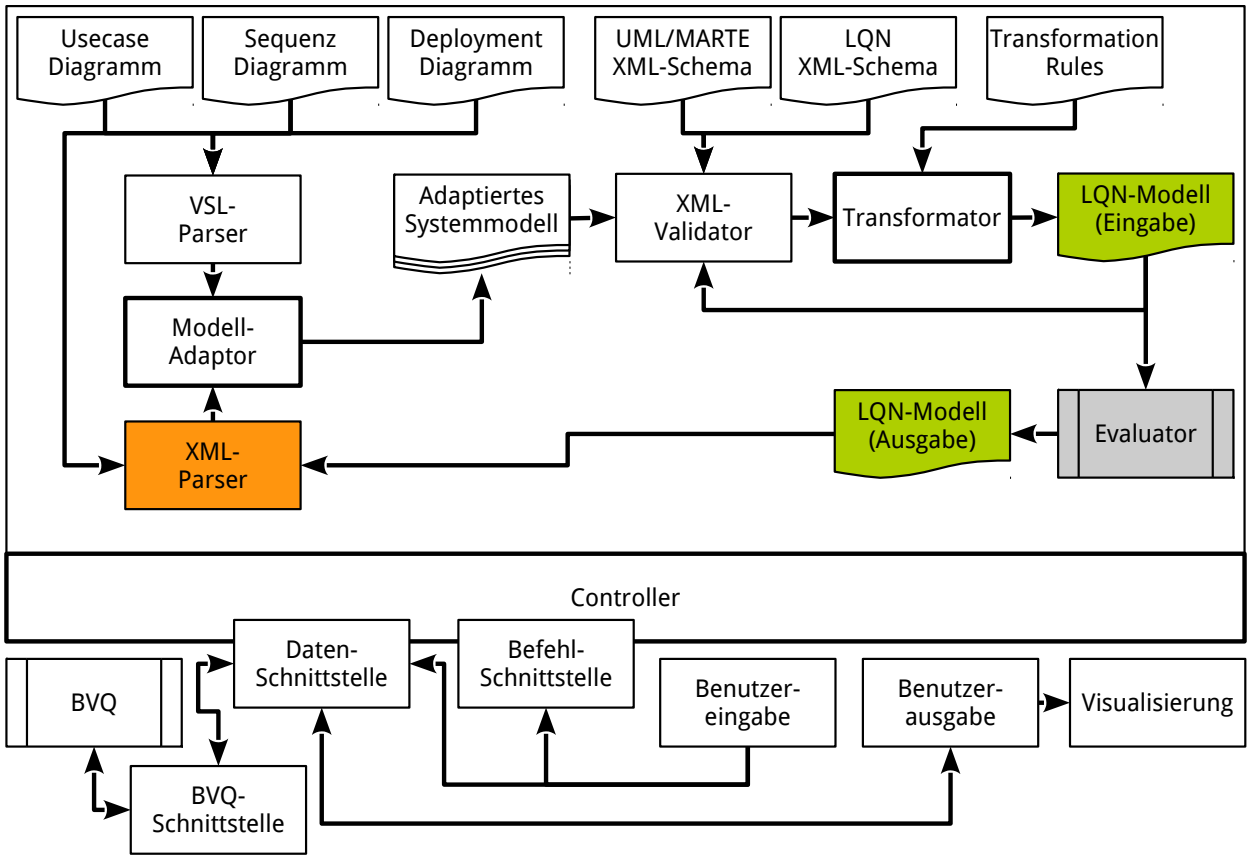
## Transformator

- Validierung der adaptierten Modelle
- Ausführung der Transformation
- Ausgabe des Leistungsmodells



## Evaluator

- Auswertung des LQN-Modells über externen Aufruf des Solvers
- Einlesen der Leistungsindizes mittels XML-Parser
- Bisher Rückführung an Datenschnittstelle (nicht Systemmodell)



## Test-Aufbau

Vergleichstest

- Internes Prognose-Tool DiskMagic als Referenz
- Bisher keine Vergleichsmessung an realen Systemen
- Angestrebte Abweichung des Modells von unter 20%

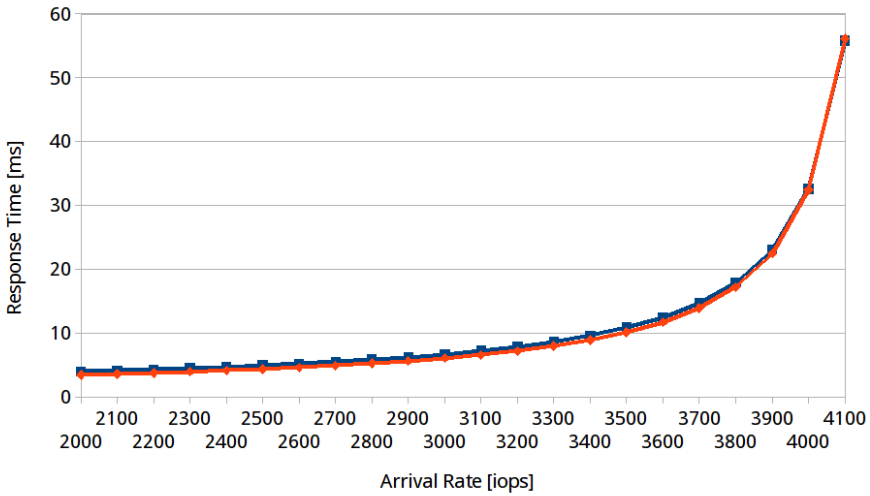
Drei Testkonfigurationen (Disk-Subsysteme)

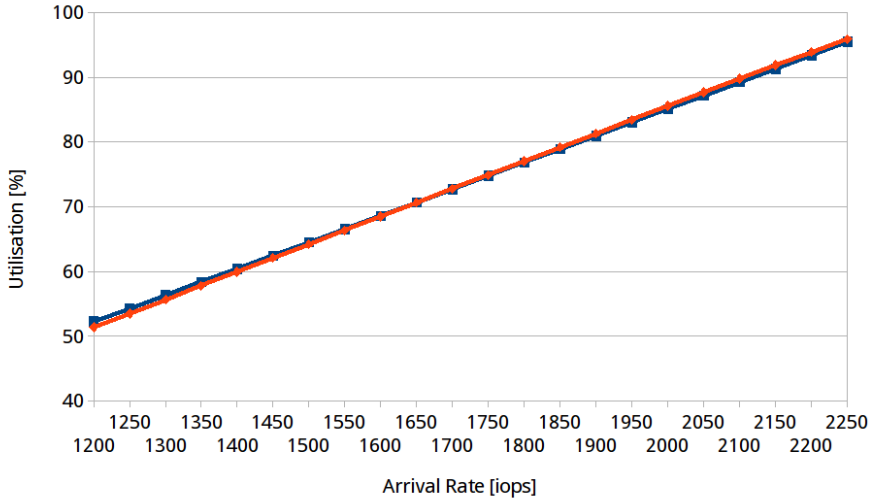
- IBM DS8870 (RAID 5, 6+1, 2.5”15k RPM)
- IBM DS8100 (RAID 5, 6+1, 3.5”10k RPM)
- IBM V7000 (RAID 5, 6+1, 2.5”15k RPM)

Sechs Test-Szenarios

- 4 homogene Zugriffsverhalten
- 1 heterogenes Zugriffsverhalten ohne Cache
- 1 heterogenes Zugriffsverhalten mit Cache

## Testergebnisse DS8870 Test 1 sequentielles Lesen

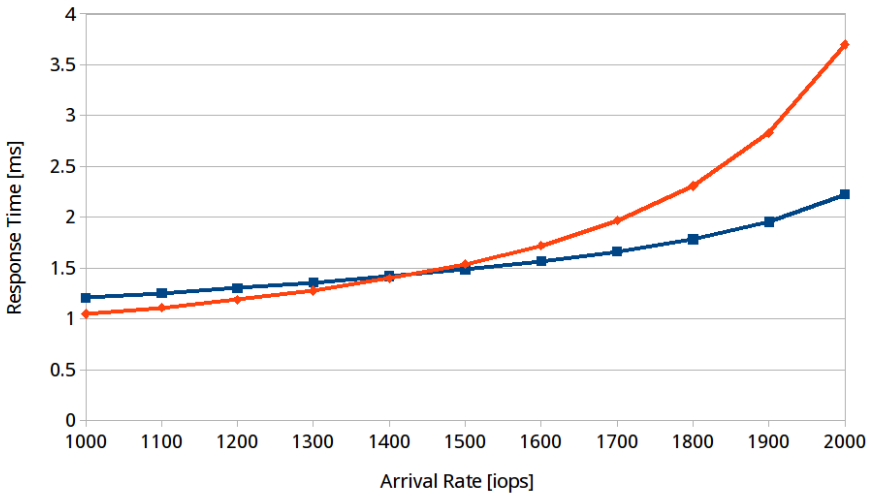


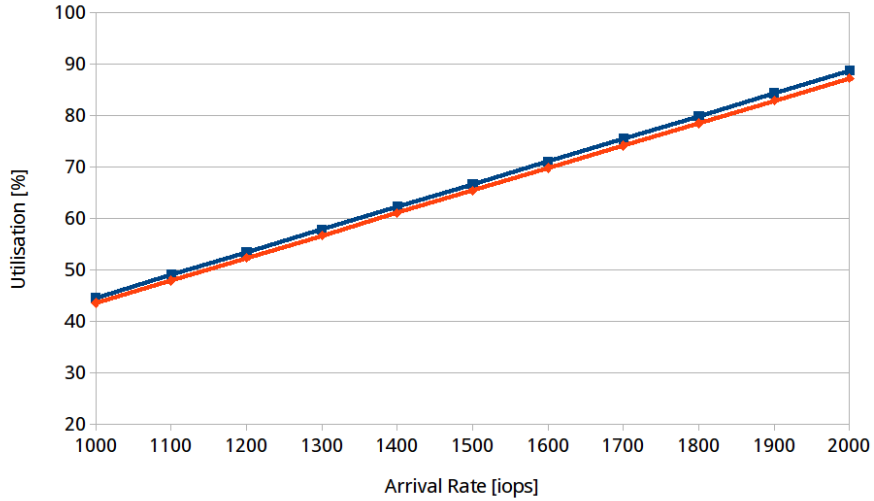




Testergebnis DS8870 Test 1 ■:DiskMagic, ♦:Modell

| Ankunftsrate | 2000 | 2500 | 3000 | 3500 | 4000 |
|---|---|---|---|---|---|
| Antwortzeit DM | 4.06 | 5.01 | 6.76 | 10.91 | 32.62 |
| Antwortzeit Modell | 3.51 | 4.44 | 6.13 | 10.18 | 32.42 |
| Differenz | 0.55 | 0.57 | 0.63 | 0.73 | 0.20 |
| Abweichung | 13% | 11% | 9% | 7% | 1% |
| Auslastung HDD DM | 48.77 | 60.10 | 71.48 | 82.95 | 94.48 |
| Auslastung HDD Modell | 47.50 | 59.38 | 71.24 | 83.10 | 94.83 |
| Differenz | 1.27 | 0.72 | 0.24 | -0.14 | -0.34 |
| Abweichung | 3% | 1% | 1% | -1% | -1% |

## Testergebnisse DS8870 Test 6 heterogener Zugriff mit Cache

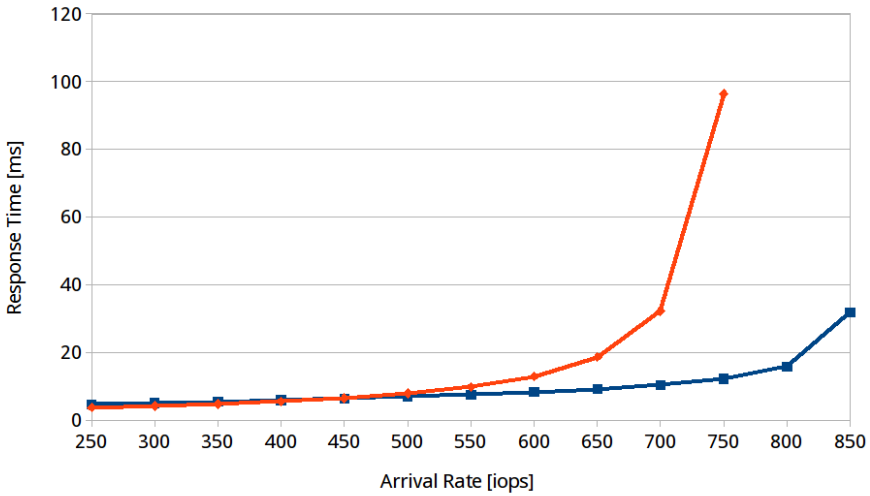


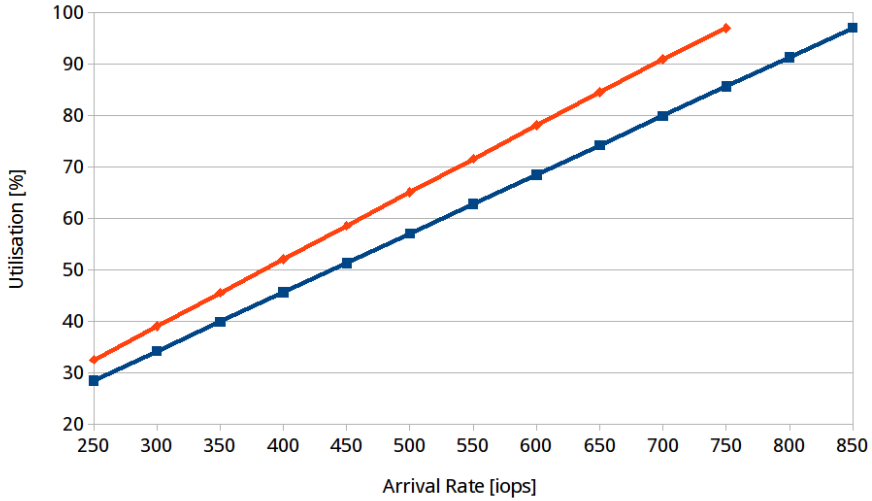




Testergebnis DS8870 Test 6 ■:DiskMagic, ♦:Modell

| Ankunftsrate | 1200 | 1400 | 1600 | 1800 | 2000 |
|---|---|---|---|---|---|
| Antwortzeit DM | 1.30 | 1.41 | 1.56 | 1.78 | 2.22 |
| Antwortzeit Modell | 1.19 | 1.40 | 1.72 | 2.31 | 3.70 |
| Differenz | 0.11 | 0.01 | -0.15 | -0.52 | -1.47 |
| Abweichung | 9% | 1% | -10% | -29% | -66% |
| Auslastung HDD DM | 53.44 | 62.24 | 71.06 | 79.88 | 88.70 |
| Auslastung HDD Modell | 52.33 | 61.05 | 69.77 | 78.5 | 87.22 |
| Differenz | 1.11 | 1.19 | 1.29 | 1.38 | 1.48 |
| Abweichung | 2% | 2% | 2% | 2% | 2% |

## Testergebnisse DS8100 Test 5 heterogener Zugriff ohne Cache




Testergebnis DS8100 Test 5 ■:DiskMagic, ♦:Modell

| Ankunftsrate | 300 | 400 | 500 | 600 | 700 |
|---|---|---|---|---|---|
| Antwortzeit DM | 5.21 | 6.07 | 7.10 | 8.43 | 10.50 |
| Antwortzeit Modell | 4.37 | 5.7 | 8.01 | 13.07 | 32.35 |
| Differenz | 0.84 | 0.37 | -0.90 | -4.63 | -21.84 |
| Abweichung | 16% | 6% | -12% | -54% | -207% |
| Auslastung HDD DM | 34.24 | 45.66 | 57.07 | 68.49 | 79.90 |
| Auslastung HDD Modell | 39.06 | 52.08 | 65.09 | 78.07 | 90.95 |
| Differenz | -4.81 | -6.41 | -8.01 | -9.57 | -11.04 |
| Abweichung | -14% | -14% | -14% | -14% | -14% |

## Stand des Arbeit, Probleme und Ausblick

Stand des Arbeit

- Modellierung und SPE-Prozess prototypisch implementiert
- Automatisierung Werkzeuge fortgeschritten
- Akzeptable Testergebnisse für Teilkonfigurationen (Referenz DM)

Probleme

- DiskMagic als Referenz möglicherweise unzulänglich
- Reifegrad Tools (Papyrus/QVT-o)
- Lizenzfrage LQN-Solver offen

Ausblick

- Vergleichstests mit realen Messergebnissen
- Verfeinerung der Modelle
- Einbettung in BVQ

Vielen Dank!

## Quellen I

[Aiz+05] N. Aizikowitz u. a. “Component-Based Performance Modeling of a Storage Area Network”. In: *Proc. Winter Simul. Conf. 2005.* (2005), S. 2417–2426.

[Fra12] Greg Franks. *Layered Queueing Network Solver and Simulator User Manual*. 2012.

[Hub+10] Nikolaus Huber u. a. “Performance modeling in industry: A Case Study on Storage Virtualization Nikolaus”. In: *Proc. 32nd ACM/IEEE Int. Conf. Softw. Eng. - ICSE ’10* (2010).

[Jai91] R. K. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1991.

[RW94] C. Ruemmler und J. Wilkes. “An introduction to disk drive modeling”. In: *Computer (Long. Beach. Calif).* (1994), S. 17–28.

## Quellen II

[Smi93] Connie U. Smith. *Performance Evaluation of Computer and Communication Systems*. Lecture Notes in Computer Science. Springer, 1993.

[Tat+10] Jon Tate u. a. *Implementing the IBM System Storage SAN Volume Controller V5.1*. IBM Redbook, 2010.

[TMW10] Mirco Tribastone, Philip Mayer und Martin Wirsing. "Performance Prediction of Service-Oriented Systems with Layered Queueing Networks". In: *Leveraging Appl. Form. Methods, Verif. Valid.* (2010), S. 51–65.

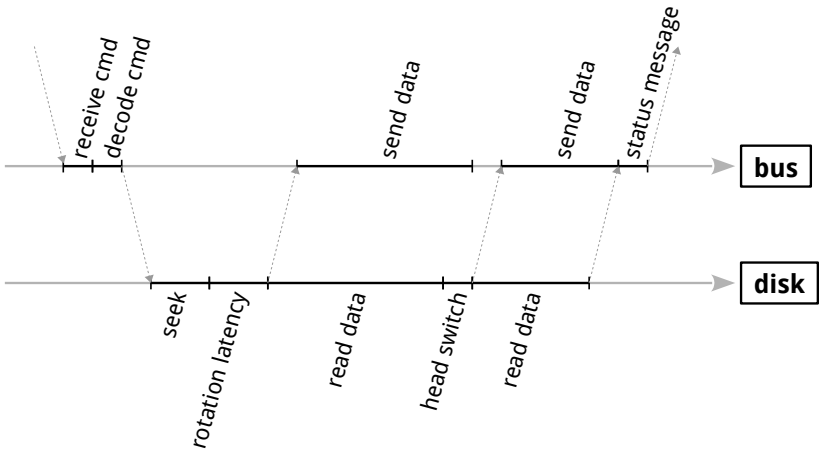
[Vog13] Benjamin Vogler. *Modellierung und Performance-Analyse einer verteilten Storageinfrastruktur*. https://wwwvs.cs.hs-rm.de/vs-wiki/index.php/MP-SS13-01. 2013.

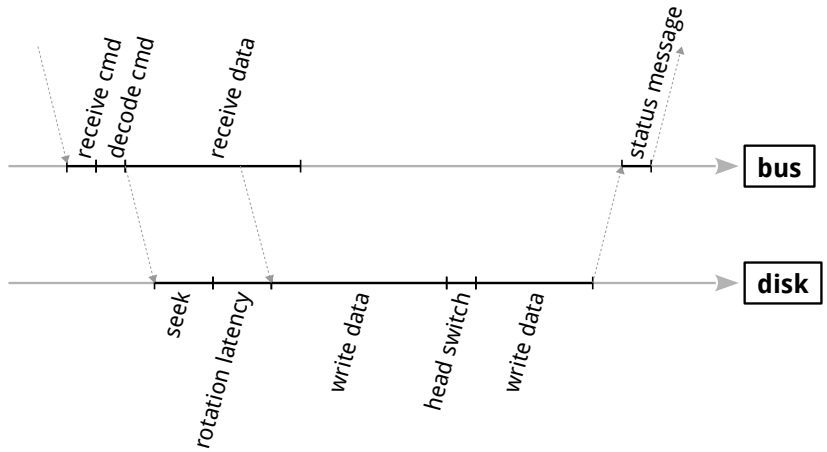
## Quellen III

[Why12] Barry Whyte. *Configuring IBM Storwize V7000 and SVC for Optimal Performance*. https://www.ibm.com/developerworks/community/blogs/storagevirtualization/entry/configuring_ibm_storwize_v7000_and_svc_for_optimal_performance_part_121?lang=en. 2012.

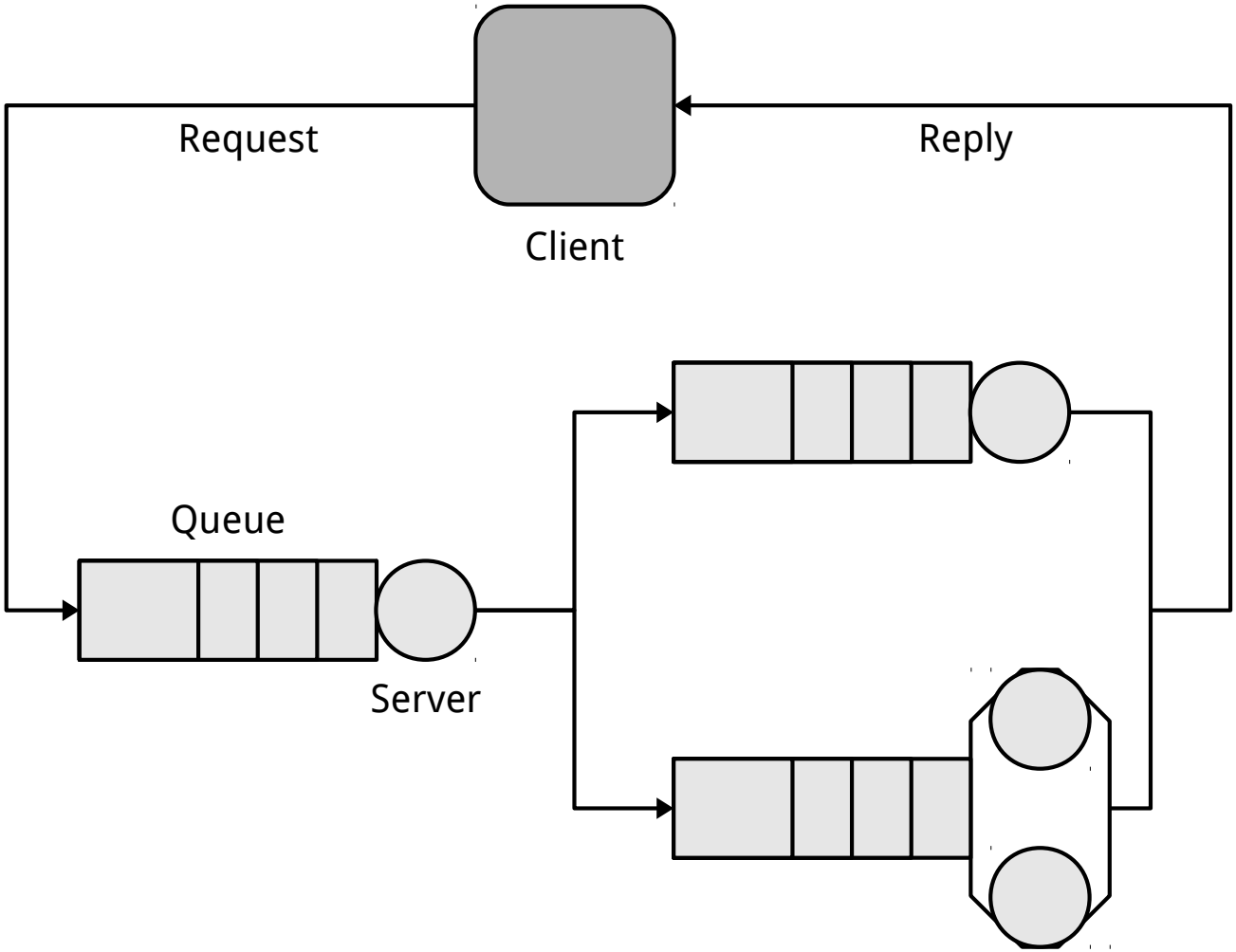
## Auflistung der Dienste (HDD)

| **RAID Level** | Read seq | Read rnd | Write seq | Write rnd |
|---|---|---|---|---|
| RAID 0 | n*r | n*r | n*w | n*w |
| RAID 5 | n*r | n*r | n*w+1*w | 2n*r + 2n*r |
| RAID 6 | n*r | n*r | n*w+2*w | 3n*r + 3n*w |
| RAID 1/10 | n*r | n*r | 2n*w | 2n*w |



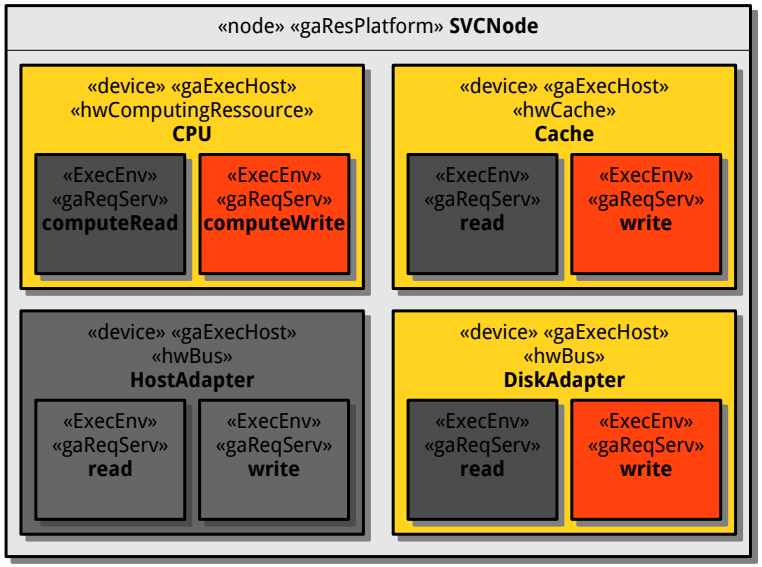



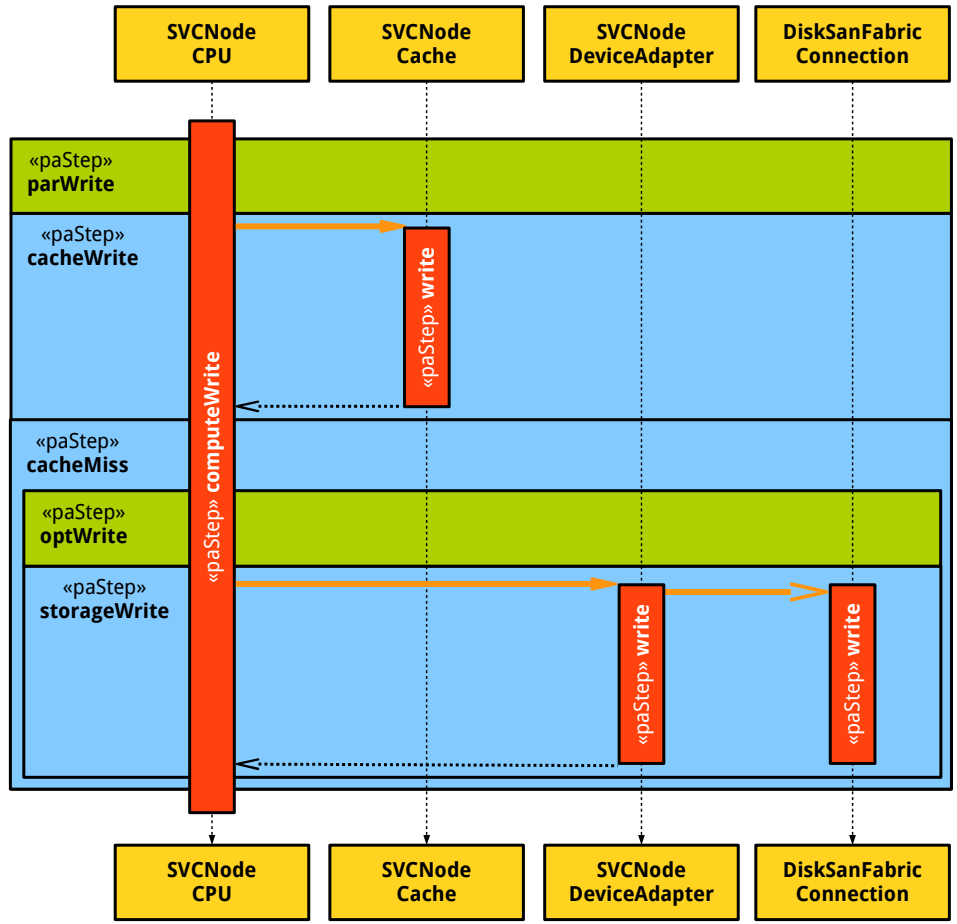
## Warteschlangennetzwerke



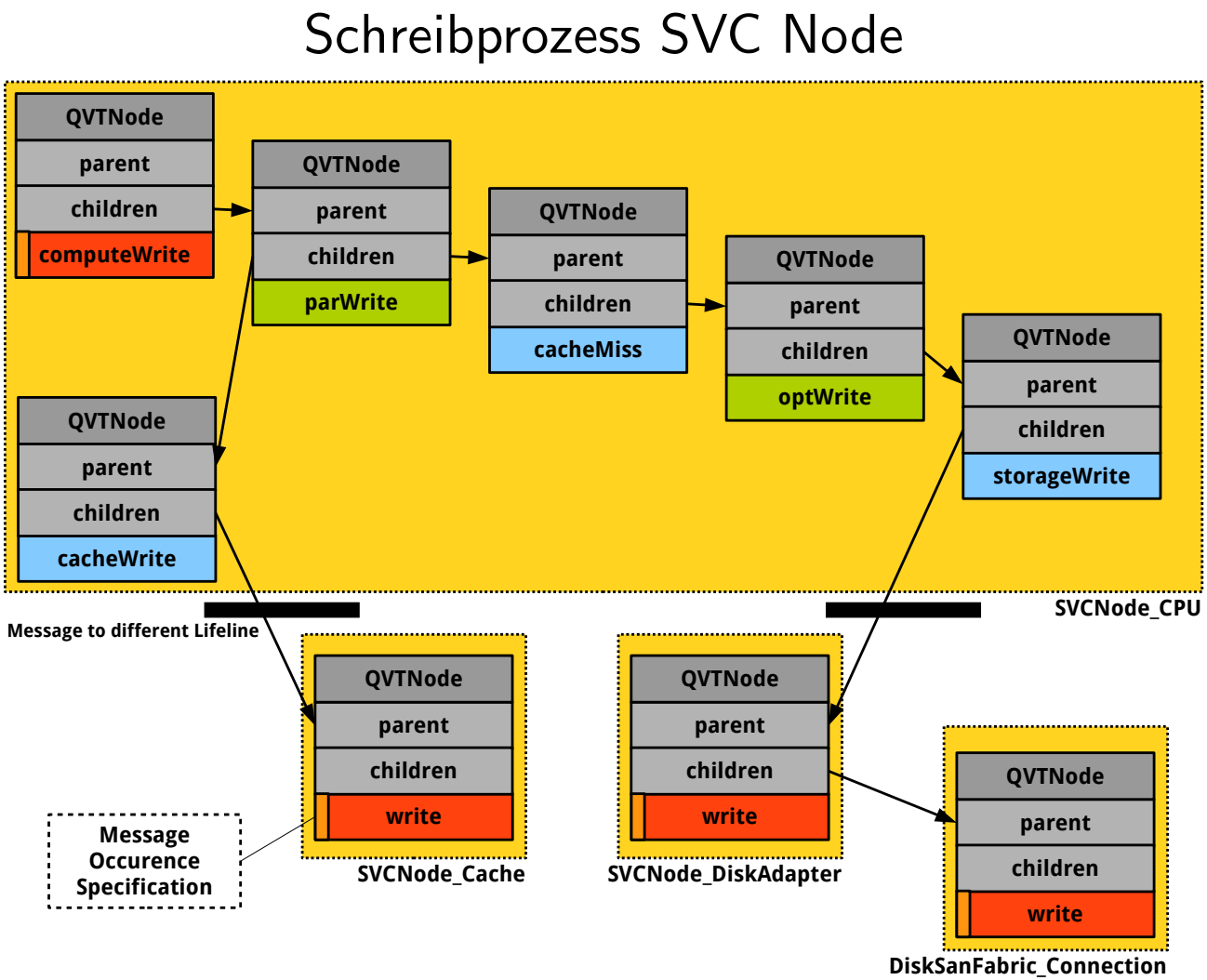
## Beispiel Systemmodell

### Schreibprozess SVC Node

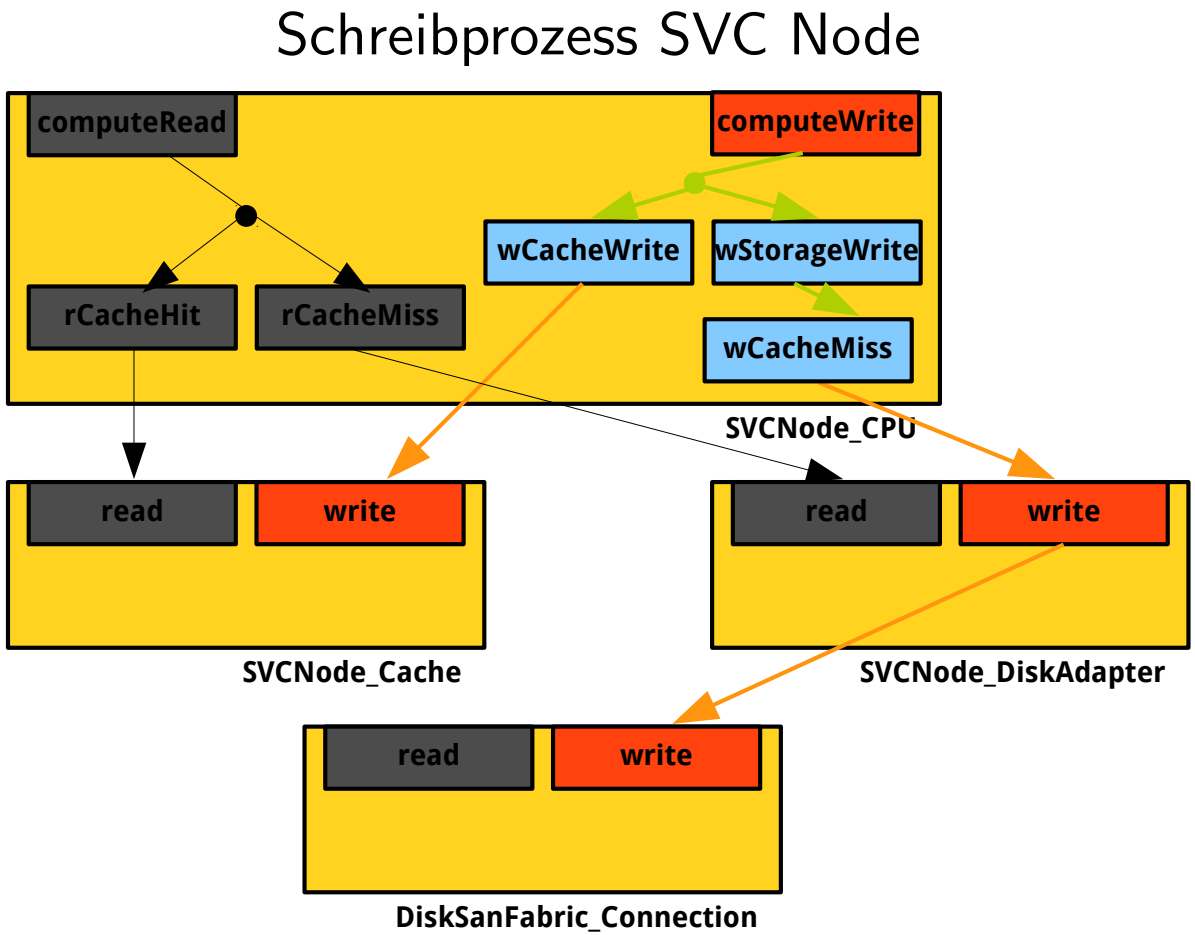


## Beispiel Transformation Zwischenmodell



## Beispiel Leistungsmodell Zwischenmodell



## Test-Konfiguration

| Systemkomponente | Eigenschaft | Wert DS8870 | Wert DS8100 |
|---|---|---|---|
| Host SAN Fabric / Adapter | Bandbreite | 4 GBit/s | 2 GBit/s |
| Disk SAN Fabric / Adapter | Bandbreite | 8 GBit/s | 4 GBit/s |
| SAN Fabrics / Adapter | Multiplizität | 2 | 2 |
| CPUs | Abarbeitungsrate | $10^7$ IOPS | $10^7$ IOPS |
| CPU (SVC Node) | Multiplizität | 2 | 2 |
| CPU (Disk-Subsystem) | Multiplizität | 2 | 2 |
| Cache (SVC Node) | Transfer Rate | 30 GB/s | 30 GB/s |
| Cache (SVC Node) | Multiplizität | 2 | 2 |
| RAID Controller | Level | 5 | 5 |
| RAID Controller | Stripe Width | 6 | 6 |
| HDD | Multiplizität | 7 | 7 |
| HDD | RPM | 15k | 10k |
| HDD | Bandbreite | 125-175 MB/s (300-500 IOPS) | 100-150 MB/s (200-300 IOPS) |

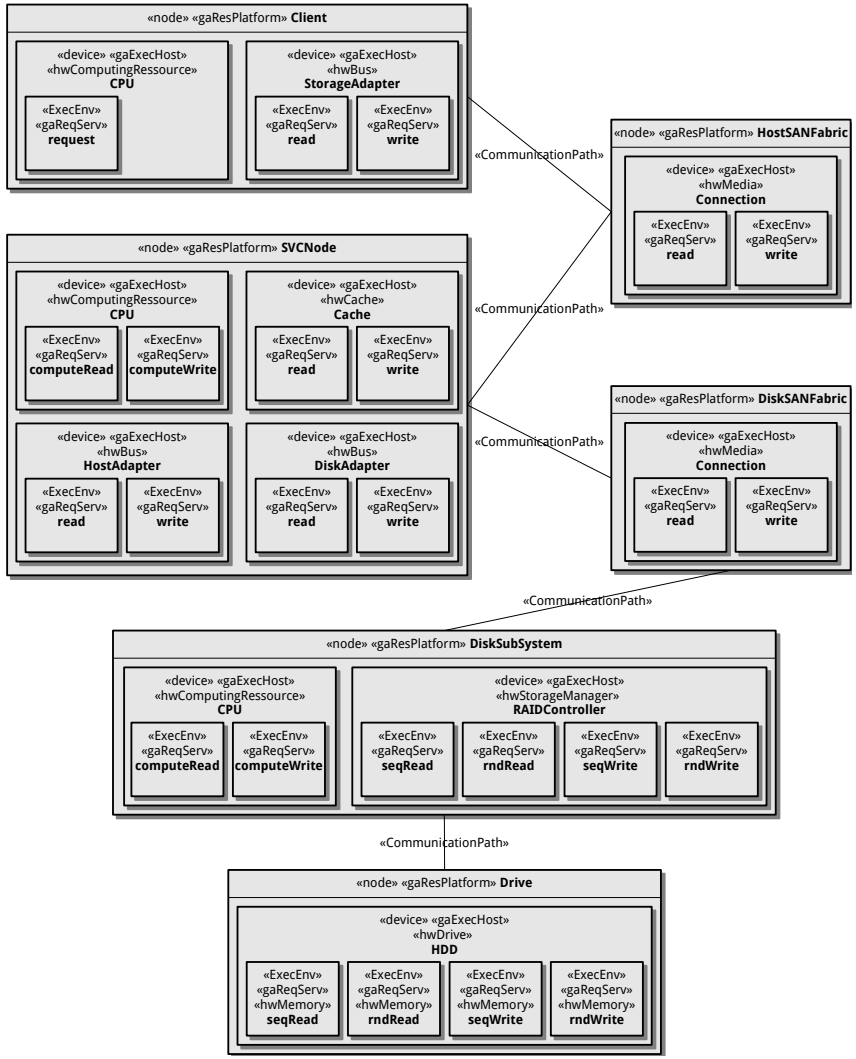
## Test-Szenarios

| Parameter | Test 1 | Test 5 | Test 6 |
|---|---|---|---|
| Ankunftsrate | | variabel | |
| Lesen (%) | 100 | 50 | 35 |
| Schreiben (%) | 0 | 50 | 65 |
| Lesen Cache Hit (%) | 0 | 0 | 50 |
| Lesen Cache Miss (%) | 100 | 100 | 50 |
| Schreiben Cache Hit (%) | 0 | 0 | 20 |
| Schreiben Cache Miss (%) | 100 | 100 | 80 |
| Lesen sequentiell (%) | 100 | 50 | 30 |
| Lesen zufällig (%) | 0 | 50 | 70 |
| Schreiben sequentiell (%) | 0 | 50 | 20 |
| Schreiben zufällig (%) | 0 | 50 | 80 |
| Transfergröße | 8KB | 8KB | 8KB |

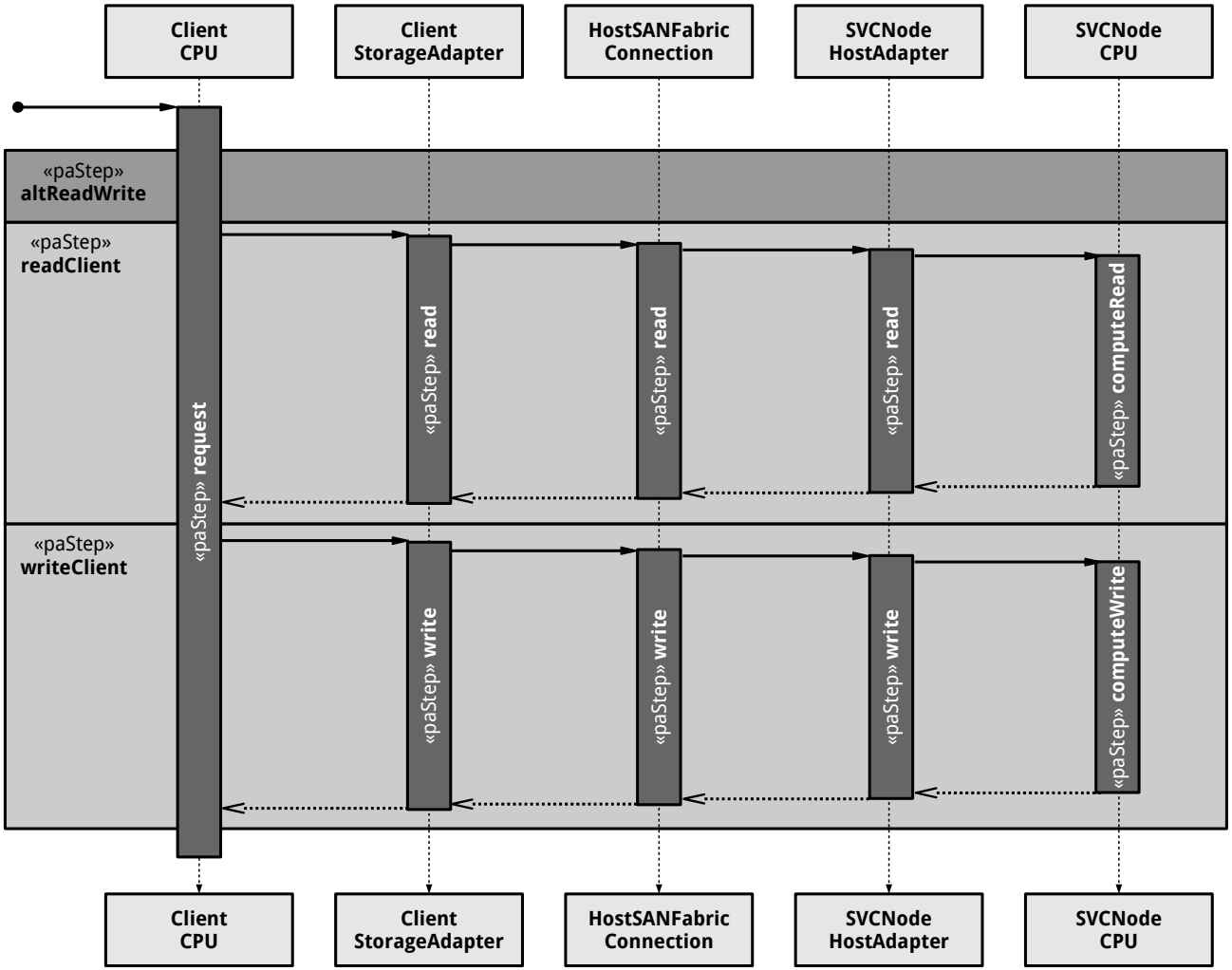
## Test-Szenarios

| Parameter | Test 1 | Test 5 | Test 6 |
|---|---|---|---|
| Ankunftsrate | | variabel | |
| Lesen (%) | 100 | 50 | 35 |
| Schreiben (%) | 0 | 50 | 65 |
| Lesen Cache Hit (%) | 0 | 0 | 50 |
| Lesen Cache Miss (%) | 100 | 100 | 50 |
| Schreiben Cache Hit (%) | 0 | 0 | 20 |
| Schreiben Cache Miss (%) | 100 | 100 | 80 |
| Lesen sequentiell (%) | 100 | 50 | 30 |
| Lesen zufällig (%) | 0 | 50 | 70 |
| Schreiben sequentiell (%) | 0 | 50 | 20 |
| Schreiben zufällig (%) | 0 | 50 | 80 |
| Transfergröße | 8KB | 8KB | 8KB |

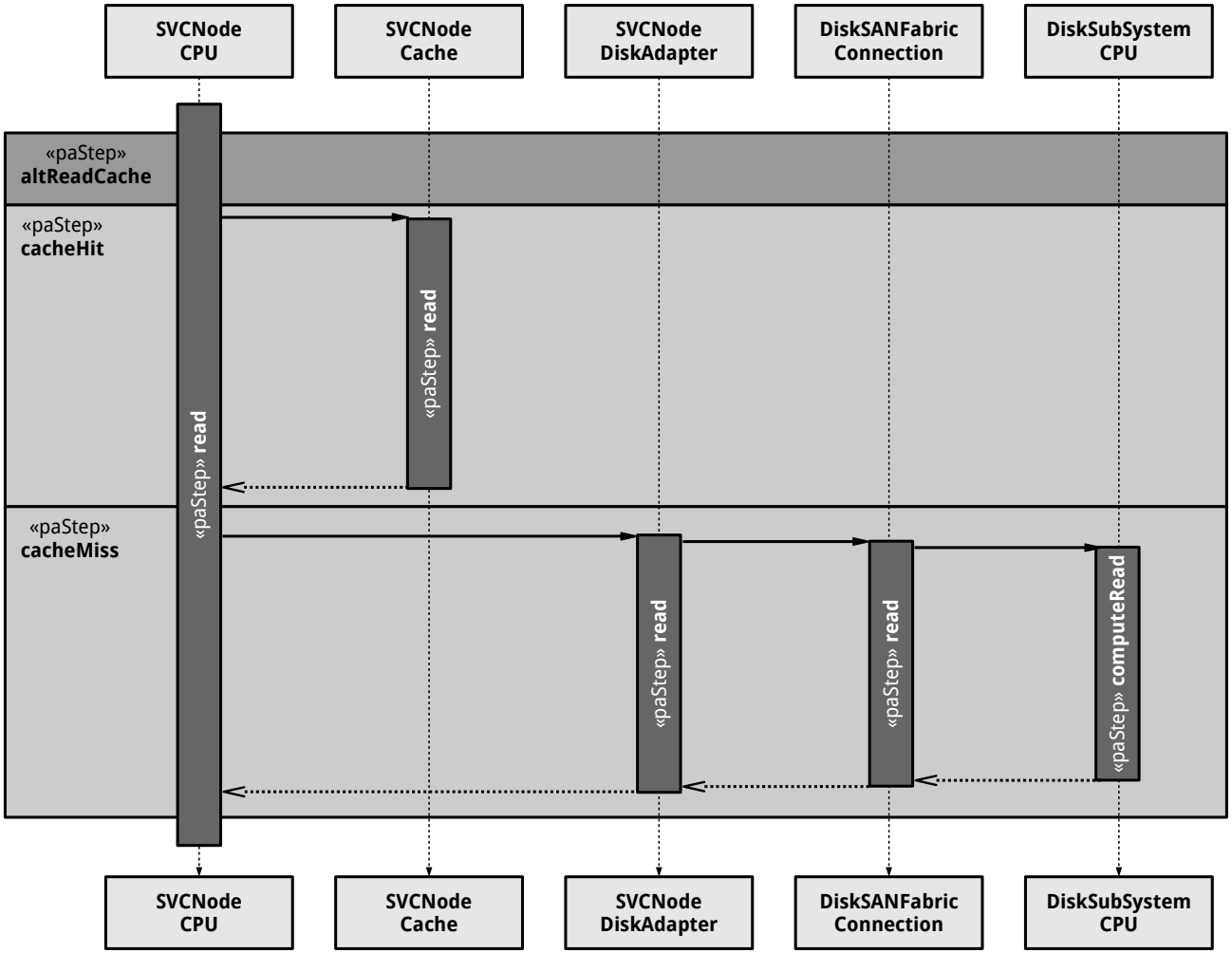
## Deployment Diagramm



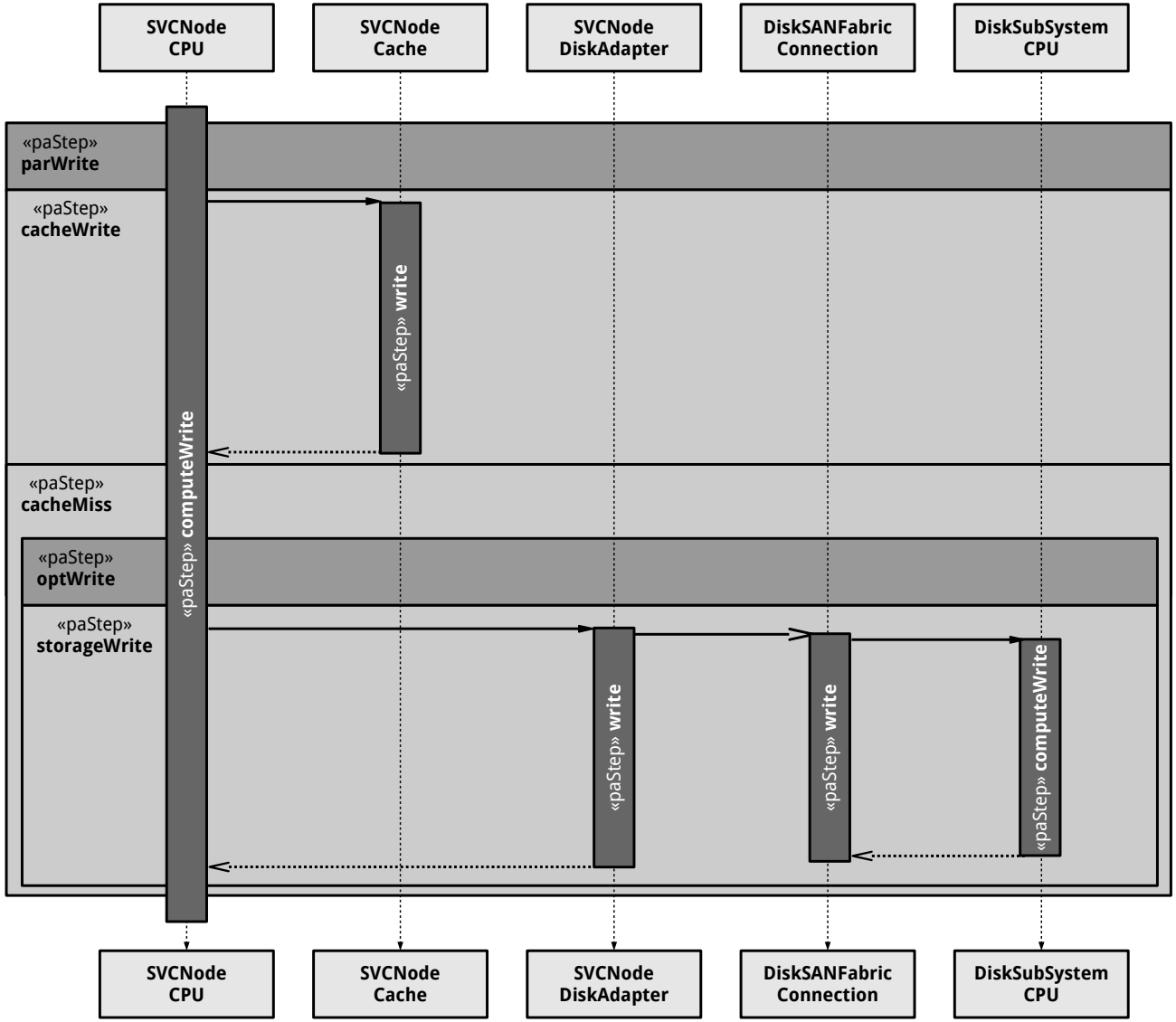
## Sequenz Diagramm



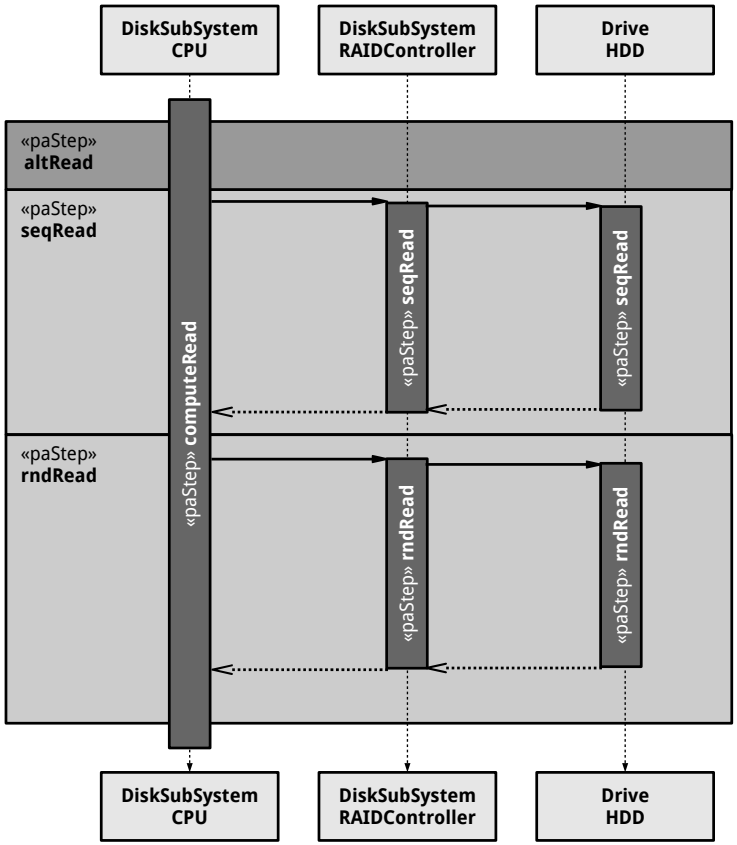
## Sequenz Diagramm



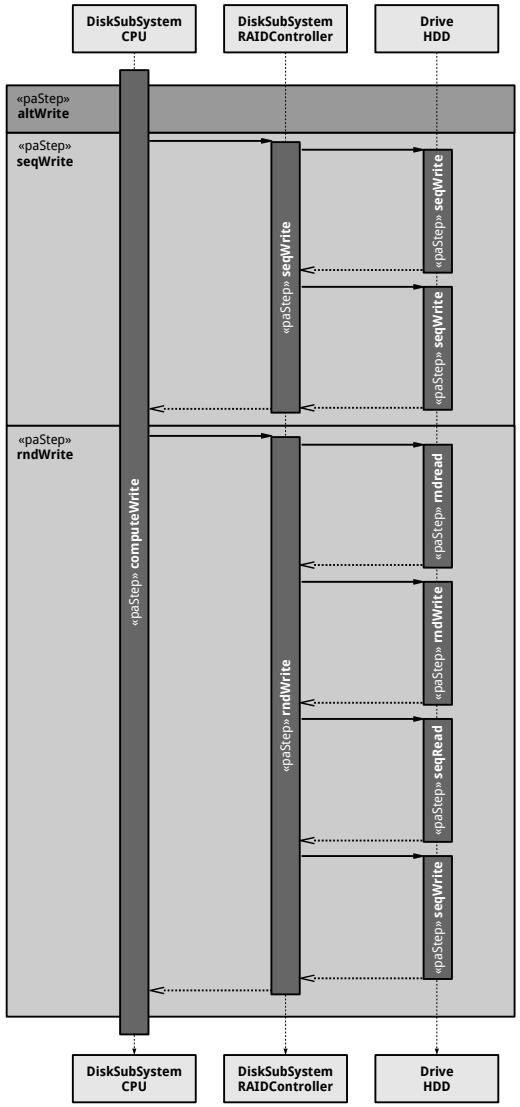
## Sequenz Diagramm



## Sequenz Diagramm



## Sequenz Diagramm



## Sequenz Diagramm