Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Descartes
research

The Descartes Modeling Language for Self-Aware Performance and Resource Management

Samuel Kounev, Fabian Brosig and Nikolaus Huber

University of Würzburg

http://se.informatik.uni-wuerzburg.de/

Software Engineering & Management 2015, Dresden, March 19, 2015

References

- S. Kounev, F. Brosig, N. Huber, and X. Zhu. **Model-Based Approach to Designing Self-Aware IT Systems and Infrastructures**. Under review. IEEE Computer Special Issue on Self-Aware and Self-Expressive Computing Systems, 2015. *Available on request.*
- S. Kounev, F. Brosig, and N. Huber. **The Descartes Modeling Language**. Technical report, Department of Computer Science, University of Wuerzburg, October 2014. [http | http | .pdf]
- F. Brosig, N. Huber, and S. Kounev. **Architecture-Level Software Performance Abstractions for Online Performance Prediction**. *Elsevier Science of Computer Programming Journal (SciCo)*, Vol. 90, Part B:71-92, 2014, Elsevier. [DOI | http | .pdf]
- N. Huber, A. van Hoorn, A. Koziolek, F. Brosig, and S. Kounev. **Modeling Run-Time Adaptation at the System Architecture Level in Dynamic Service-Oriented Environments**. *Service Oriented Computing and Applications Journal (SOCA)*, 8(1):73-89, 2014, Springer-Verlag. [DOI | .pdf]
- F. Brosig, P. Meier, S. Becker, A. Koziolek, H. Koziolek, and S. Kounev. **Quantitative Evaluation of Model-Driven Performance Analysis and Simulation of Component-based Architectures**. *IEEE Transactions on Software Engineering (TSE)*, 2014, IEEE, Preprint. [DOI | .pdf]
- F. Gorsler, F. Brosig, and S. Kounev. **Performance Queries for Architecture-Level Performance Models**. In *5th ACM/SPEC International Conference on Performance Engineering (ICPE 2014)*, Dublin, Ireland, 2014. ACM, New York, NY, USA. 2014. [DOI | .pdf]
- N. Herbst, N. Huber, S. Kounev and E. Amrehn. **Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning**. *Concurrency and Computation - Practice and Experience, John Wiley and Sons, Ltd.*, 26(12):2053-2078, 2014. [DOI | http | .pdf]
- S. Spinner, G. Casale, F. Brosig, S. Kounev. **Evaluating Approaches to Resource Demand Estimation**. Elsevier Performance Evaluation Journal, 2015. Under publication. Available on request.
- *N. Herbst, S. Kounev and R. Reussner.* ***Elasticity: What it is, and What it is Not.*** *In 10th Intl. Conference on Autonomic Computing (ICAC 2013), San Jose, CA, June 24-28, 2013.* [slides | http | .pdf]
- J. von Kistowski, N. Herbst, and S. Kounev. **Modeling Variations in Load Intensity over Time**. In *3rd Intl. Workshop on Large-Scale Testing (LT 2014)*, Dublin, Ireland, March 22, 2014, pages 1-4. ACM, New York, NY, USA. March 2014. [DOI | slides | http | .pdf]
- A. Milenkoski, B. Payne, N. Antunes, M. Vieira and S. Kounev. **HInjector: Injecting Hypercall Attacks for Evaluating VMI-based Intrusion Detection Systems** (Poster Paper). In *2013 Annual Computer Security Applications Conf. (ACSAC 2013)*, New Orleans, Louisiana, USA, 2013. [.pdf]

Further References

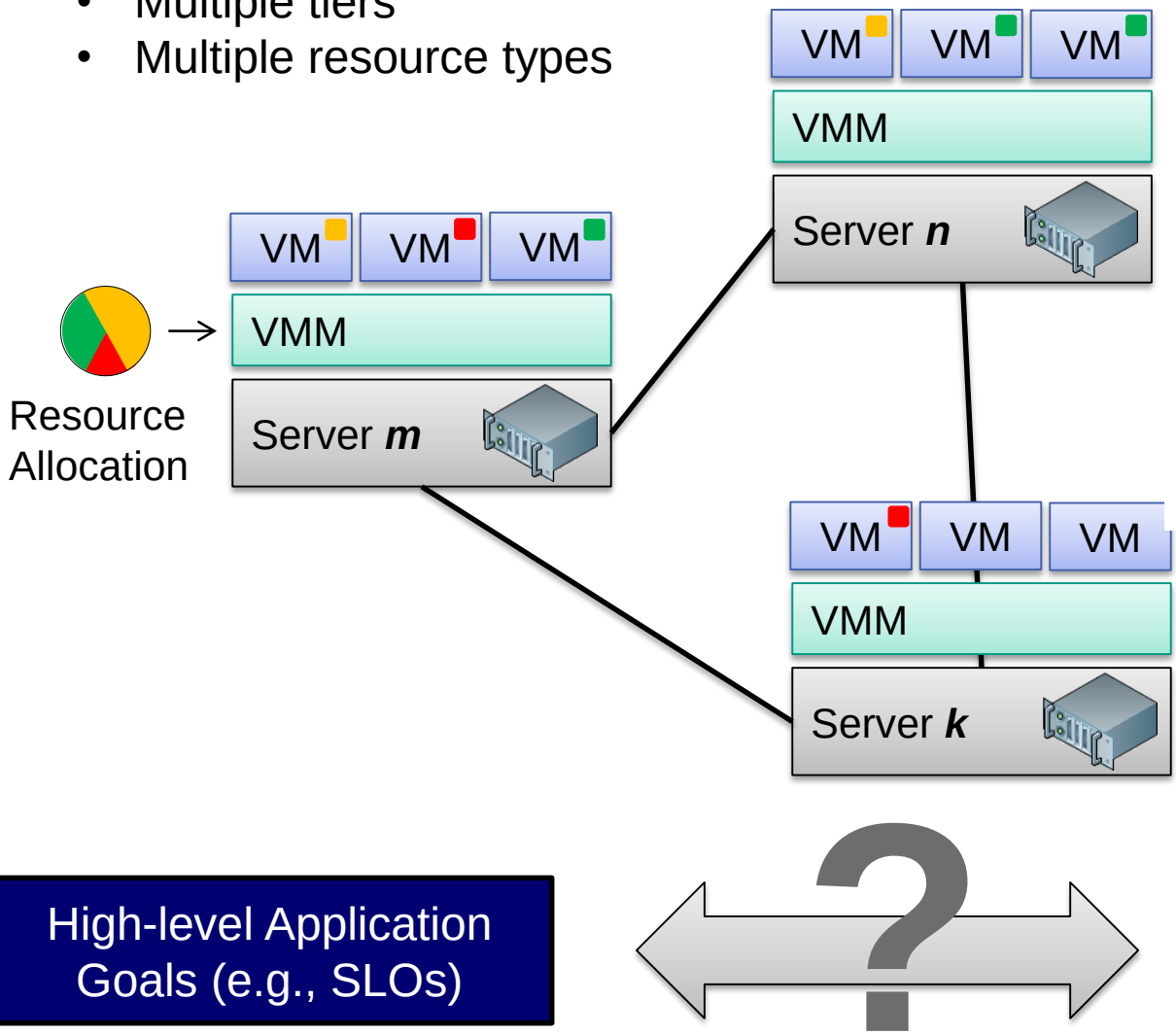
- **Fabian Brosig**. *Architecture-Level Software Performance Models for Online Performance Prediction*. PhD thesis, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, 2014. [http | http]

- **Nikolaus Huber**. *Autonomic Performance-Aware Resource Management in Dynamic IT Service Infrastructures*. PhD thesis, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, 2014. [http | http]

Motivation

Applications

- Multiple tiers
- Multiple resource types



High-level Application Goals (e.g., SLOs)

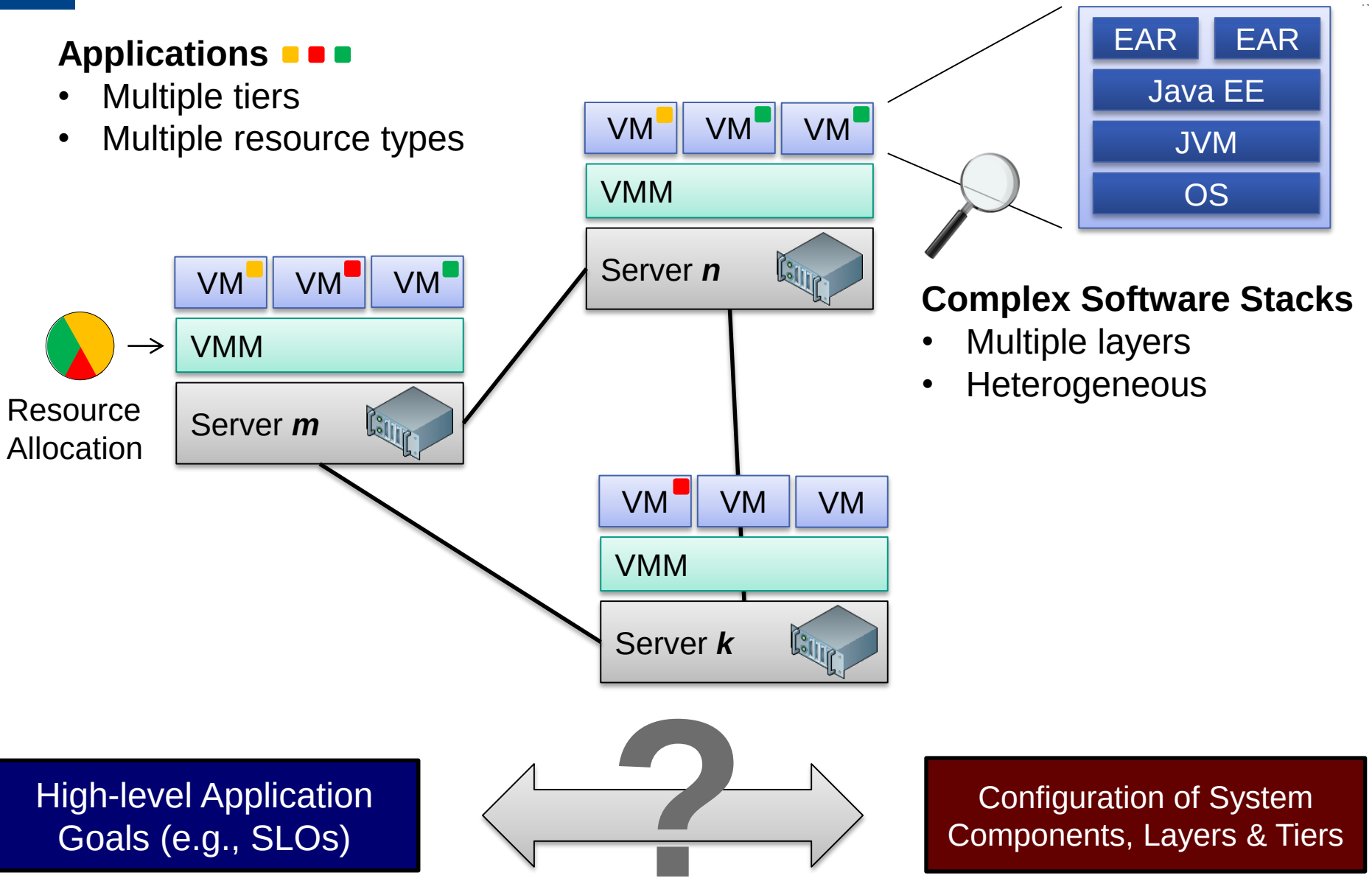
?

Configuration of System Components, Layers & Tiers

Motivation

Applications

- Multiple tiers
- Multiple resource types



Complex Software Stacks

- Multiple layers
- Heterogeneous

High-level Application Goals (e.g., SLOs)

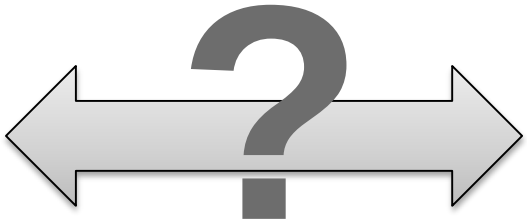
Configuration of System Components, Layers & Tiers

Semantic Gap Problem

- **Availability & Performance**
  - # requests that can be processed per sec > 1000
  - Response time of service x < 20 ms
  - Server utilization > 60% on average
  - ...
- **Efficiency**
  - Allocate only as much resources as are actually needed
- ...

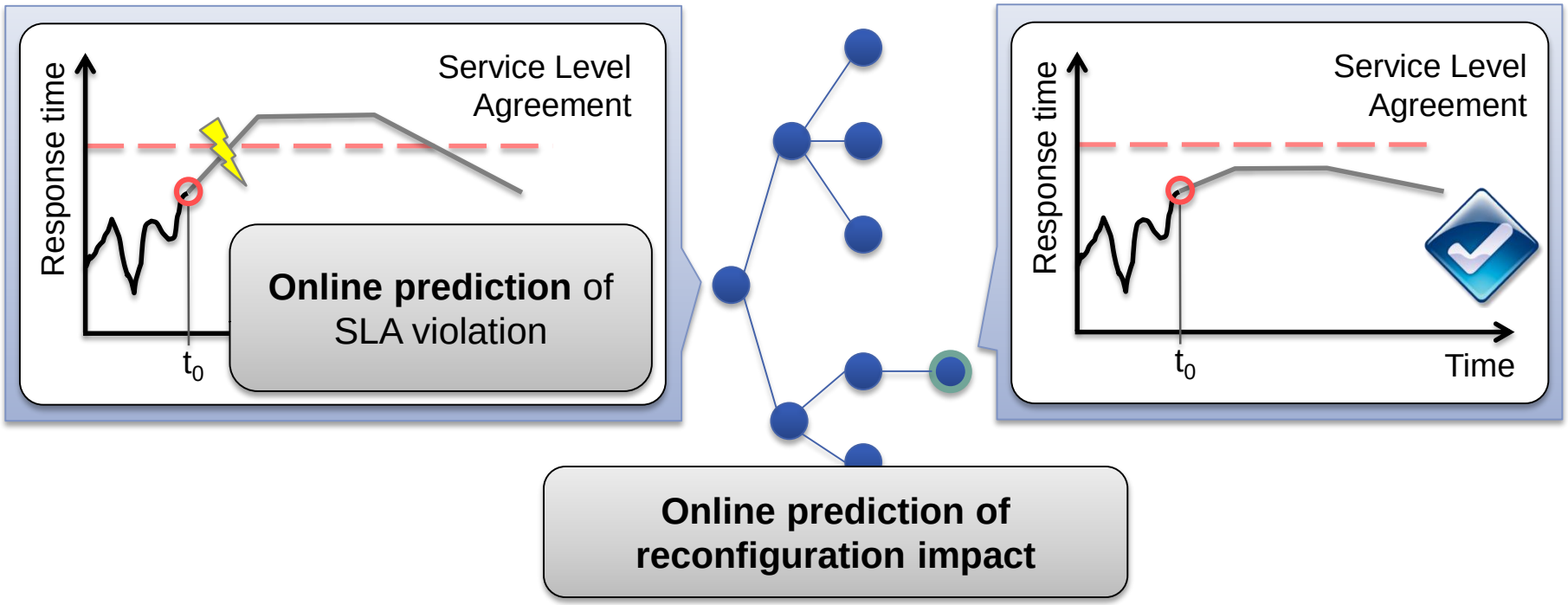
- How many vCPUs to allocate to virtual machine (VM) n?
- How much memory to allocate to VM n?
- When exactly should a reconfiguration be triggered?
- Which particular resources to scale / replicate / migrate?
- How quickly and at what granularity?

High-level Application Goals (e.g., SLOs)



Configuration of System Components, Layers & Tiers

Self-Aware Performance & Resource Management



Self-Aware Performance & Resource Management

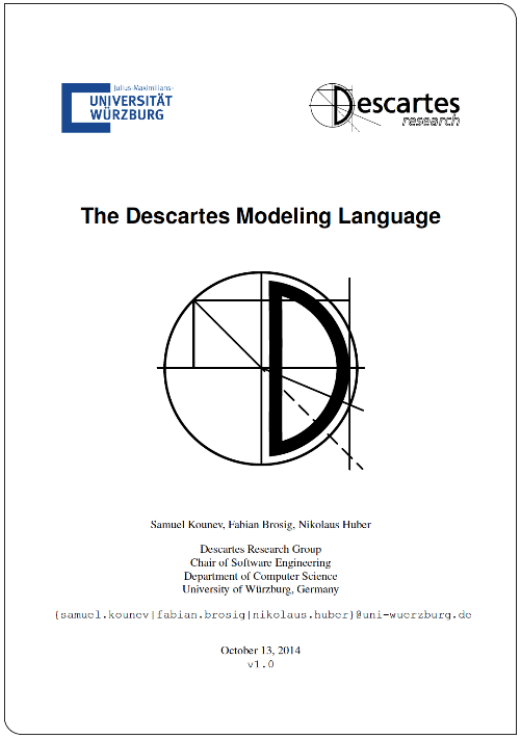
Example of Self-Aware Computing

See http://www.dagstuhl.de/15041

Dagstuhl Seminar 15041, January 18-23, 2015

Descartes Modeling Language (DML)

- Problem: How to model the performance and resource management related aspects of an IT system to enable self-aware performance and resource management?



The Descartes Modeling Language

Samuel Kounev, Fabian Brosig, Nikolaus Huber

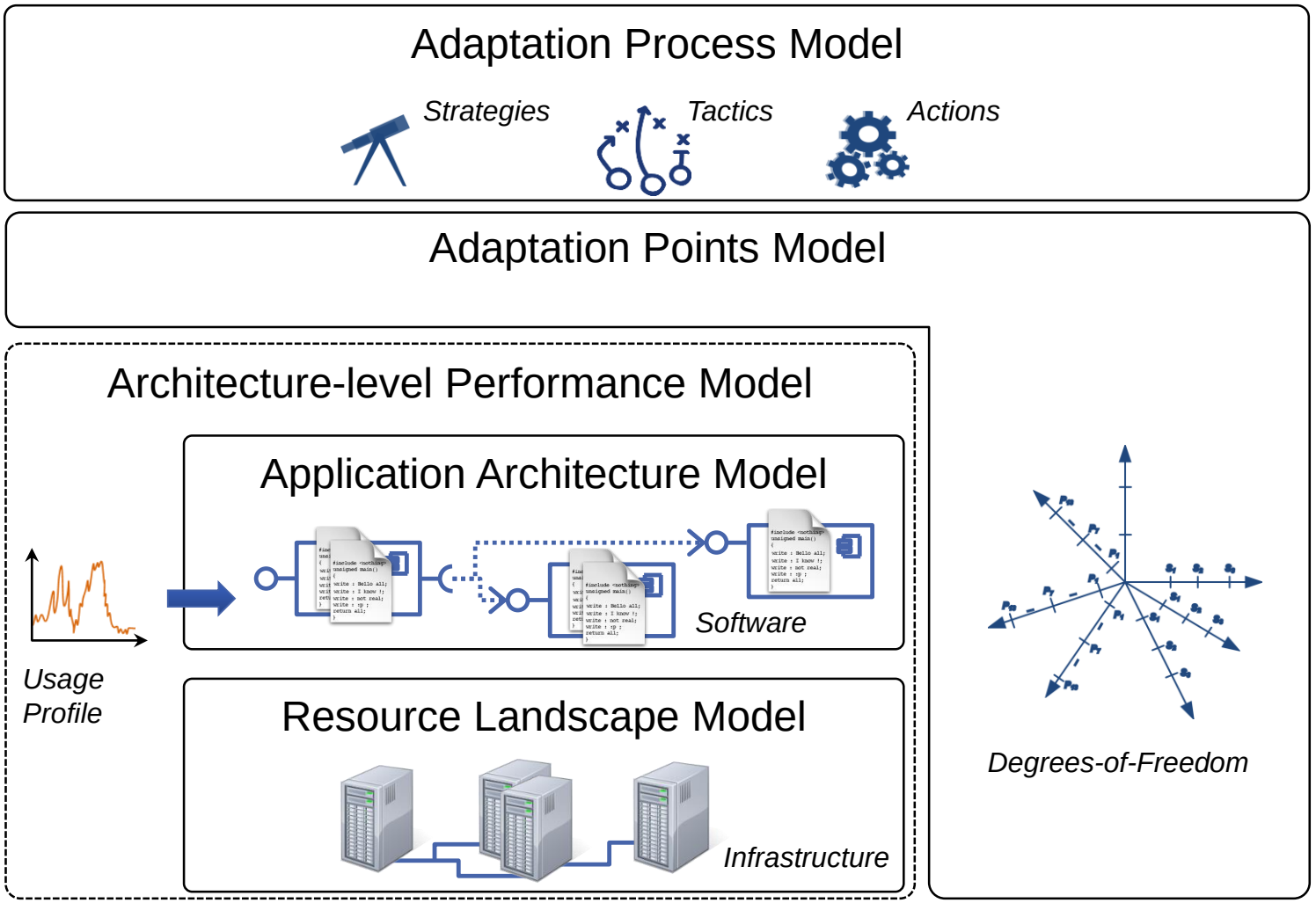
Descartes Research Group
Chair of Software Engineering
Department of Computer Science
University of Würzburg, Germany

{samuel.kounev|fabian.brosig|nikolaus.huber}@uni-wuerzburg.de

October 13, 2014
v1.0

http://descartes.tools/dml

DML Sub-Models



Big Picture

Adaptation Process Model

Strategies Tactics Actions

describes

Adaptation Process

Logical

evaluates

adapts

Adaptation Points Model

Architecture-Level Performance Model

Usage Profile Model

Application Architecture Model

A B C

<<InternalAction>> ResourceDemandX

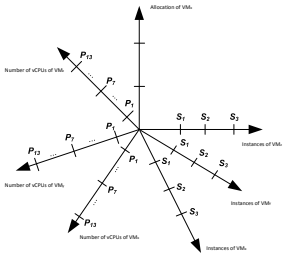
Deployment Model

Resource Landscape Model

<<Container>> Node1

<<Container>> Node2

<<Container>> Node3

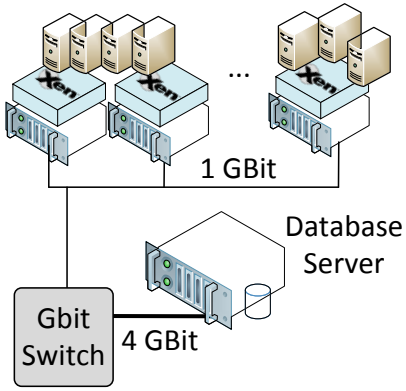


Degrees of Freedom

models

para-meterizes

Managed System

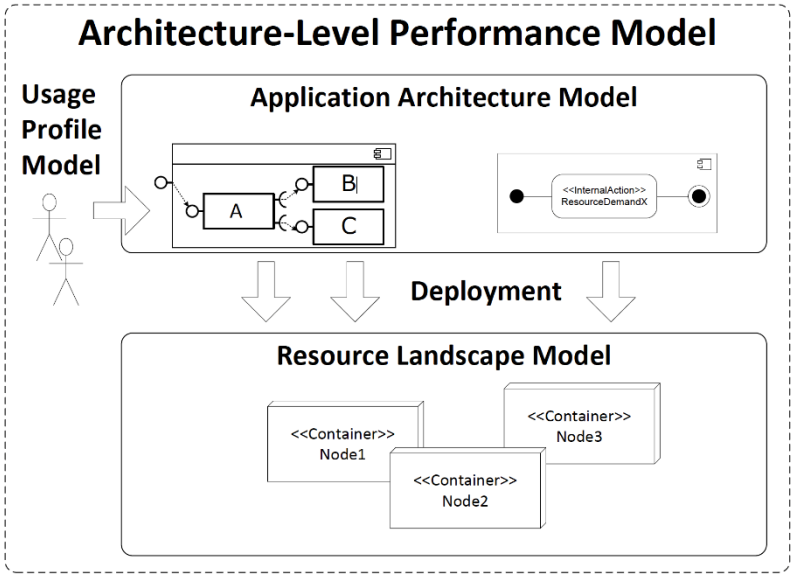


Technical

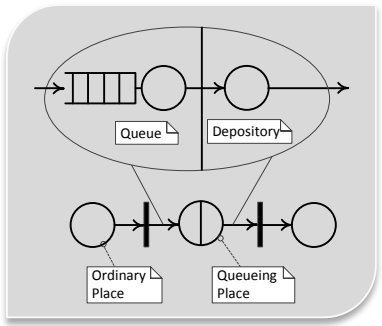
DML Instance

System

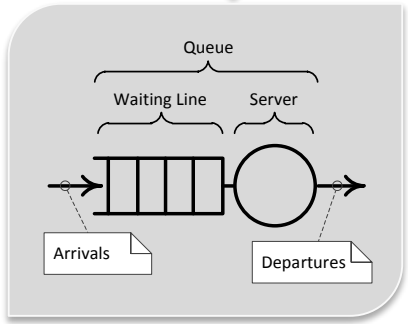
Transformations to Predictive Models



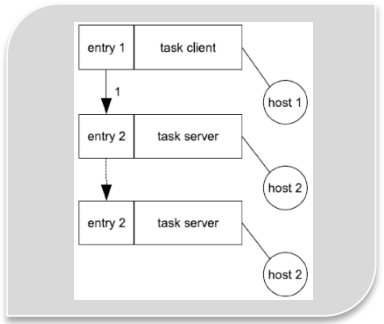
DML Instance



Queueing Petri Net



Bounds Analysis Model



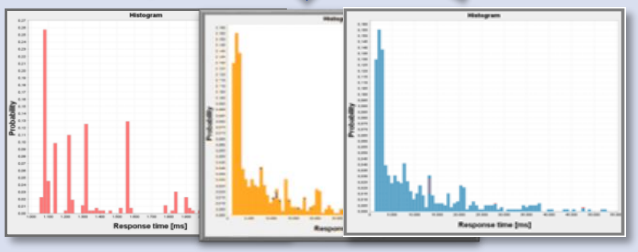
Layered Queueing Network

Tailored Model Solution

Analytical Analysis

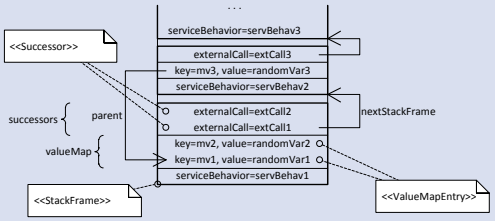
$$R \geq \max\left[N \times \max\{D_i\}, \sum_{i=1}^{K} D_i \right] \qquad X_0 \leq \min\left[\frac{1}{\max\{D_i\}}, \frac{N}{\sum_{i=1}^{K} D_i} \right]$$

$$\frac{N}{\max\{D_i\}[K+N-1]} \leq X_0 \leq \frac{N}{avg\{D_i\}[K+N-1]}$$



Analysis Results

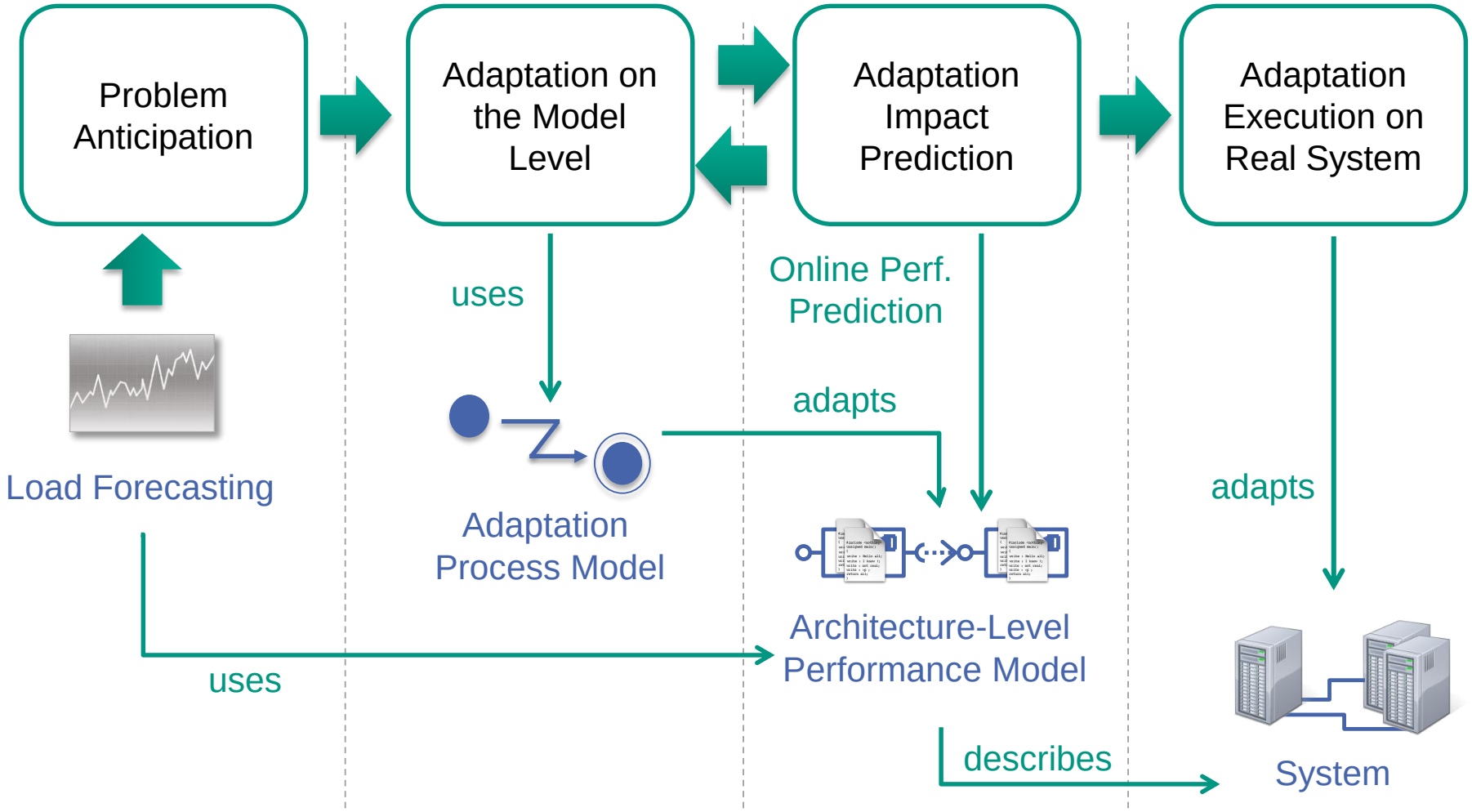
Simulative Analysis



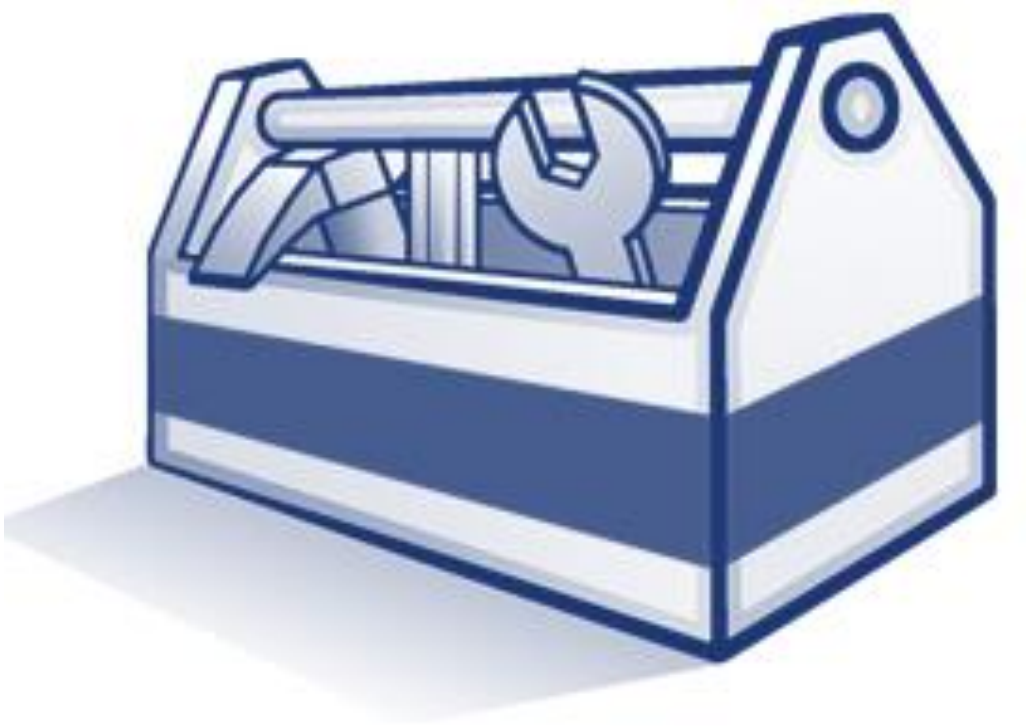
Analysis Results

F. Brosig, P. Meier, S. Becker, A. Koziolek, H. Koziolek, and S. Kounev. **Quantitative Evaluation of Model-Driven Performance Analysis and Simulation of Component-based Architectures**. *IEEE Transactions on Software Engineering (TSE)*, 2014, IEEE, Preprint. [DOI | .pdf]

Model-Based System Adaptation



Descartes Tool Chain

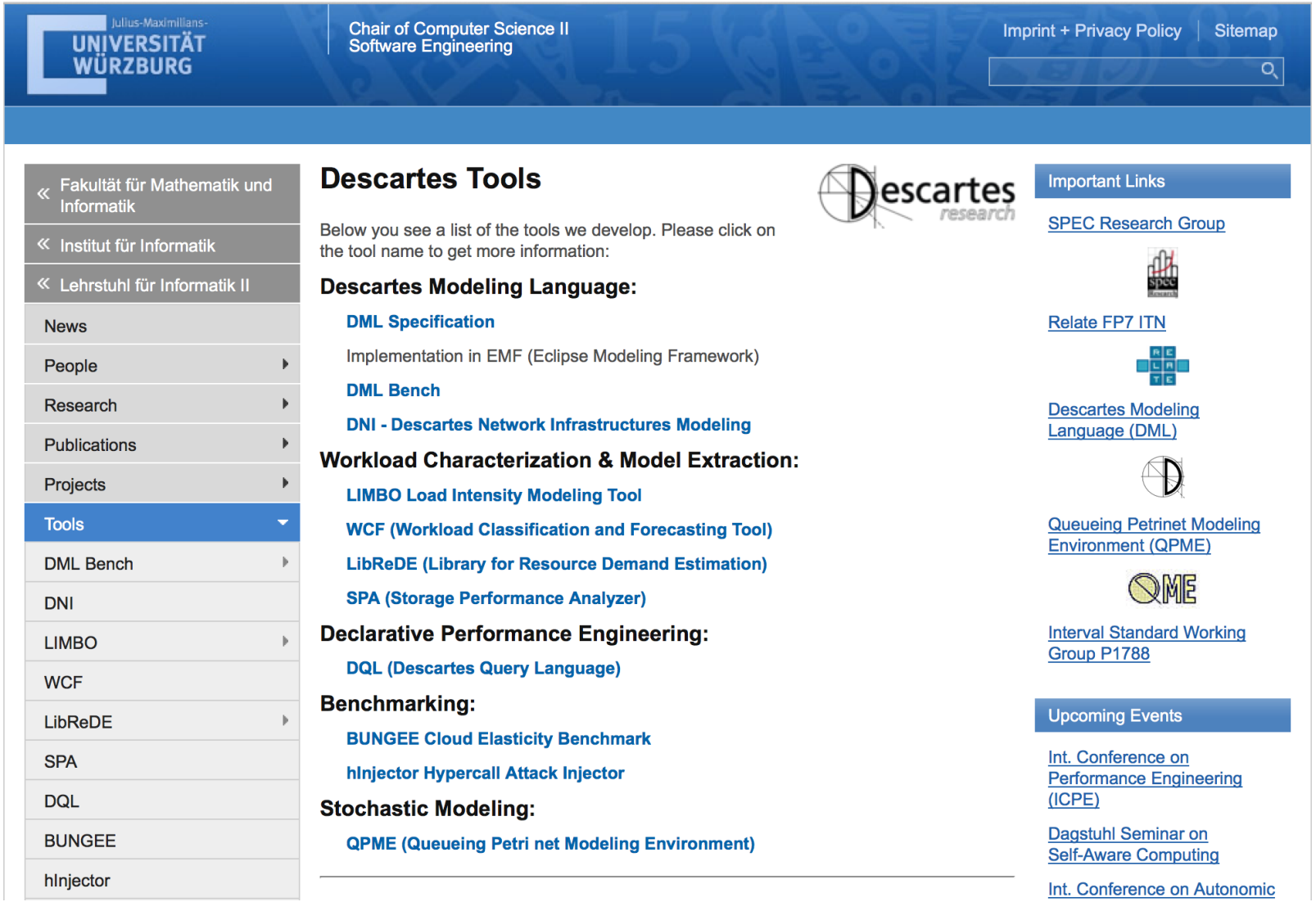


http://descartes.tools

Current Tools

- **DML** – Descartes Modeling Language (homepage, publications)
- **DML Bench** (homepage, publications)
- **DQL** – Declarative performance query language (homepage, publications)
- **LibReDE** - Library for resource demand estimation (homepage, publications)
- **LIMBO** – Load intensity modeling tool (homepage, publications)
- **WCF** – Workload classification & forecasting tool (homepage, publications)
- **BUNGEE** – Elasticity benchmarking framework (homepage, publications)
- **hInjector** – Security benchmarking tool (homepage, publications)
- **Further relevant research**
  - **http://descartes-research.net/research/research_areas/**
  - **Self Aware Computing** (publications)

http://descartes.tools



DML Bench

- **Editors**
  - Textual and graphical editors for DML models
- **Solvers**
  - Solvers for conducting performance prediction
- **S/T/A Adaptation Framework**
  - Execution of adaptation process on the model level

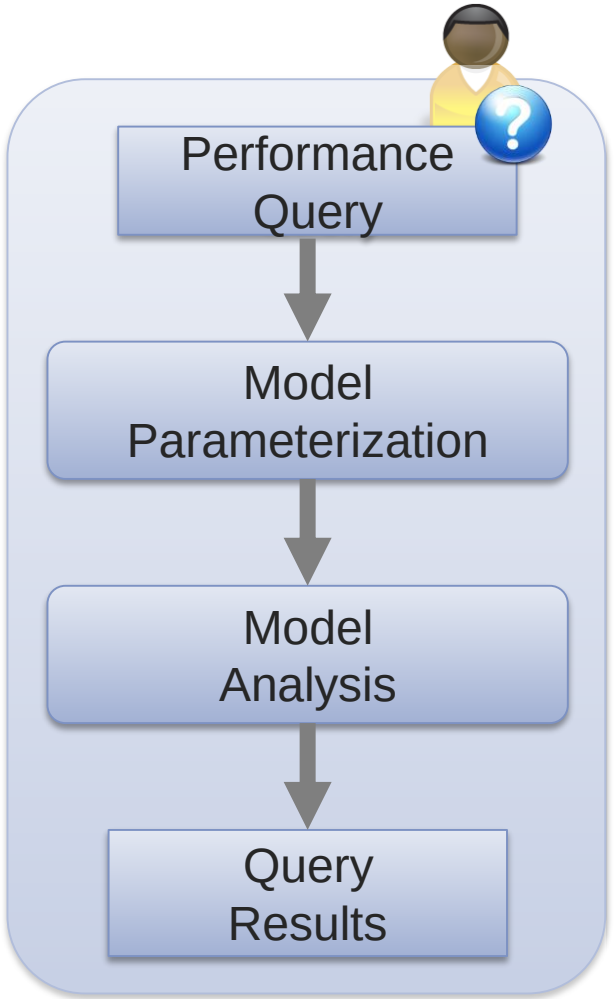
DML Bench

http://descartes.tools/dml_bench

DQL (Descartes Query Language)

Example of a
performance query specified with DQL

```
SELECT s.avgResponseTime,
app.utilization,
dbs.utilization
CONSTRAINED AS FAST
FOR RESOURCE
'ApplicationServer' AS app,
RESOURCE 'DBServer' AS dbs,
SERVICE 'processOrder' AS s;
```

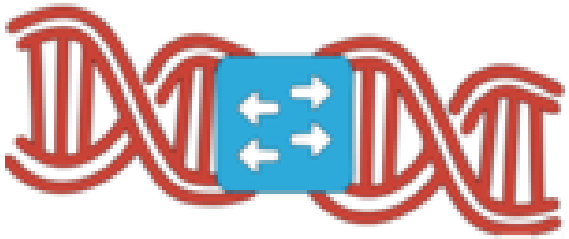


DQL

http://descartes.tools/dql

DNI - Descartes Network Infrastructure Modeling

- Language for perf. modeling of data center networks
  - network topology, switches, routers, virtual machines, network protocols, routes, flow-based configuration,...
- Model solvers based on simulation (OMNeT)



http://descartes.tools/dni

LibReDE Tool

- Problem: How to estimate the total service time of a given type of request/job at a given resource?
- **Library for Resource Demand Estimation**
  - Ready-to-use implementations of estimation approaches
  - Comparison of the accuracy of different approaches
  - Selection of a suitable approach for a given scenario

http://descartes.tools/librede

S. Spinner, G. Casale, X. Zhu, and S. Kounev. **LibReDE: A Library for Resource Demand Estimation (Demo Paper)**. In *5th ACM/SPEC Intl. Conf. on Performance Engineering (ICPE 2014)*, Dublin, Ireland, March 22-26, 2014, pages 227-228. ACM Press, New York, NY, USA. March 2014. [http | .pdf]

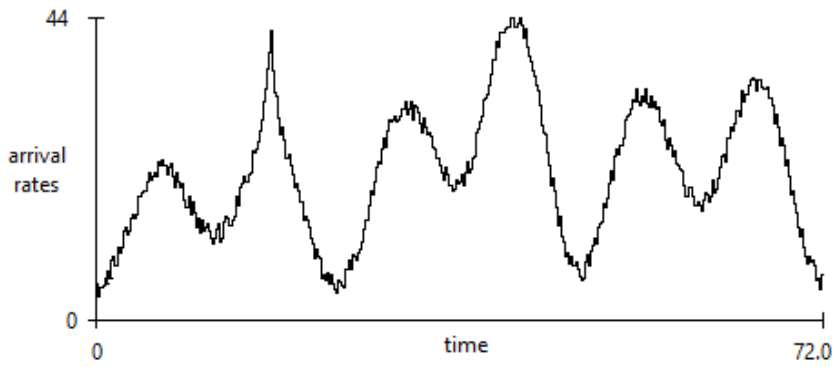
Estimation Approaches

Technique	Variant	References
Approximation with response times		Urgaonkar et al. [13] Nou et al. [14] Brosig et al. [15]
Service Demand Law		Lazowksa [4] Brosig et al. [15]
Linear regression	Least squares	Bard and Shatzoff [16] Rolia et al. [17, 18] Pacifici et al. [19] Kraft et al. [20, 21]
	Least absolute differences	Zhang et al. [22, 23, 24]
	Least trimmed squares	Casale et al. [25, 26]
Kalman filter		Zheng et al. [27, 28] Kumar et al. [29] Wang et al. [30, 31]
Optimization	Non-linear constrained optimization	Zhang et al. [32] Menascé [33]
	Quadratic programming	Liu et al. [34, 35, 36] Kumar et al. [37]
Machine learning	Clusterwise linear regression	Cremonesi et al. [38]
	Independent component analysis	Sharma et al. [39]
	Support vector machine	Kalbasi et al. [40]
	Pattern matching	Cremonesi et al. [41, 42]
Maximum likelihood estimation		Kraft et al. [20] Perez et al. [21]
Gibbs sampling		Sutton and Jordan [43] Wang et al. [44]
Demand Estimation with Confidence (DEC)		Kalbasi et al. [45, 46]

S. Spinner, G. Casale, F. Brosig, S. Kounev. **Evaluating Approaches to Resource Demand Estimation**. Elsevier Performance Evaluation Journal. Under publication. Available on request.

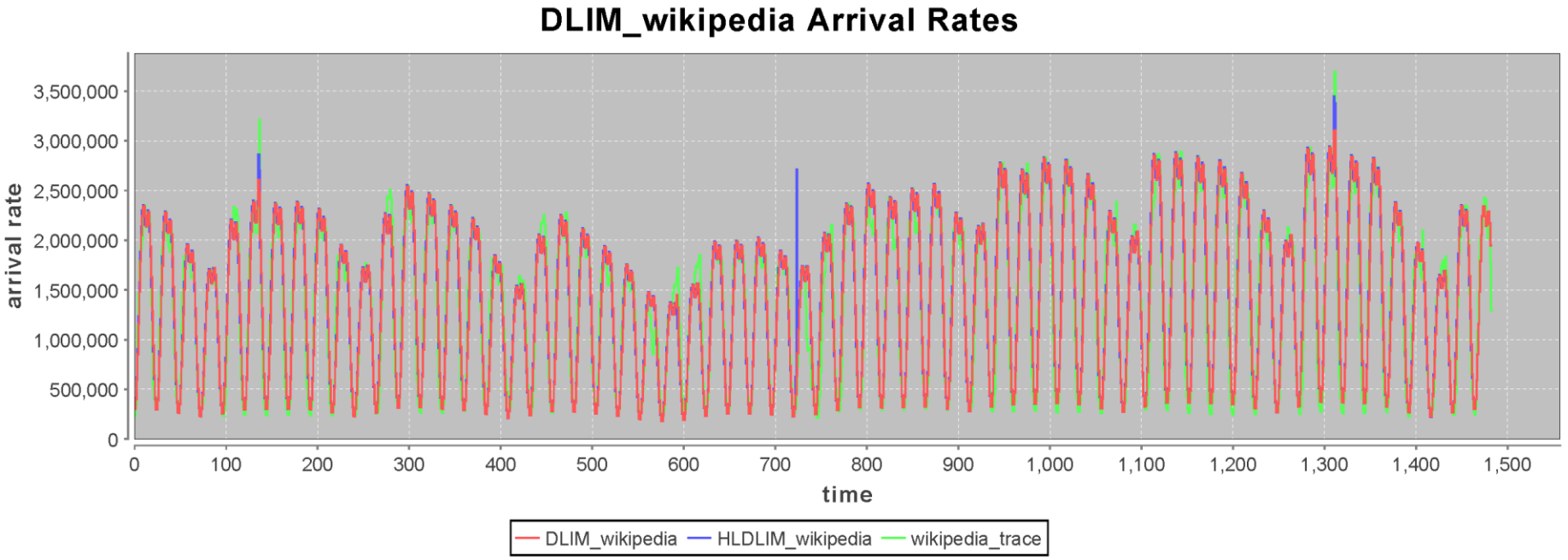
LIMBO Tool

- Problem:
  - How to capture the load intensity variations (e.g. requests per sec) in a compact mathematical model?
  - How to forecast the load intensity (requests per sec) in future time horizons?
- **Load Intensity Modeling & Forecasting Tool**

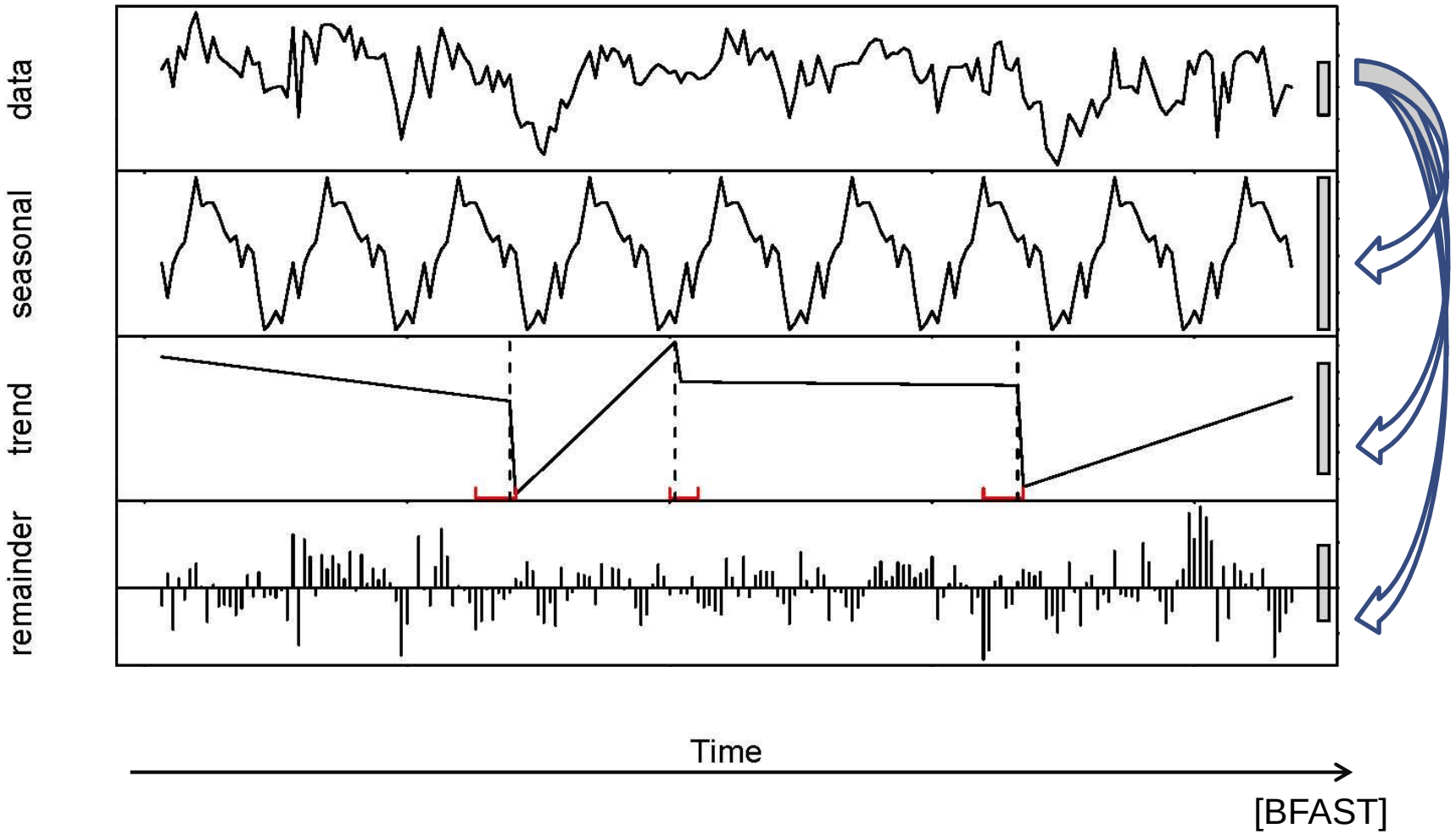


http://descartes.tools/limbo

Example: Wikipedia Workload



Time Series Analysis



LIMBO Tool

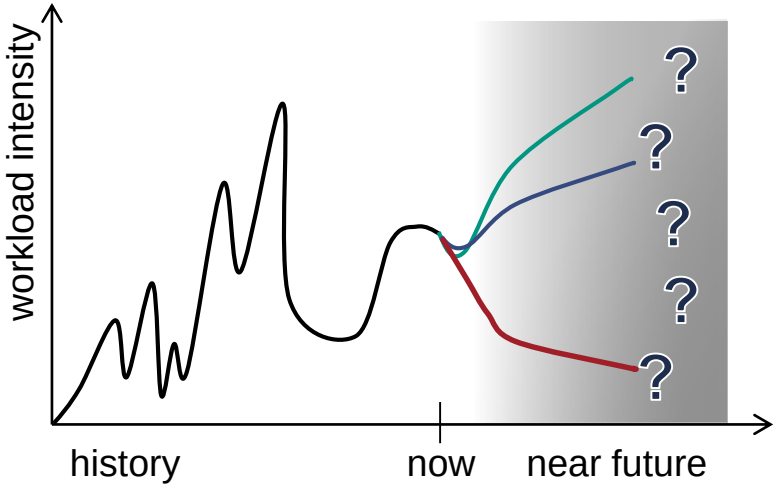
http://descartes.tools/limbo

J. von Kistowski, N. Herbst, and S. Kounev. **LIMBO: A Tool For Modeling Variable Load Intensities (Demo Paper)**. In *5th ACM/SPEC International Conference on Performance Engineering (ICPE 2014)*, Dublin, Ireland, March 22-26, 2014, ICPE '14, pages 225-226. ACM, New York, NY, USA. March 2014. [DOI | slides | http | .pdf]

J. von Kistowski, N. Herbst, and S. Kounev. **Modeling Variations in Load Intensity over Time**. In *3rd Intl. Workshop on Large-Scale Testing (LT 2014)*, Dublin, Ireland, March 22, 2014, pages 1-4. ACM, New York, NY, USA. March 2014. [DOI | slides | http | .pdf]

LIMBO Tool

- **Workload Classification & Forecasting (WCF)**
  - Use of multiple alternative forecasting methods in parallel
  - Selection of method based on its accuracy in the past



http://descartes.tools/wcf

Systems Benchmarking

Metrics and benchmarks for quantitative evaluation of

1. Resource elasticity
2. Performance isolation
3. Intrusion detection (and prevention)
4. ...

[geek & poke]

S. Kounev. **Quantitative Evaluation of Service Dependability in Shared Execution Environments** (Keynote Talk). In 11th Intl. Conf. on Quantitative Evaluation of SysTems (QEST 2014), Florence, Italy, September 8-12, 2014. [slides | extended abstract]

Elasticity

Def: The degree to which a system is able to **adapt** to **workload changes** by **provisioning and deprovisioning** resources in an **autonomic manner**, such that at each point in time the **available resources match** the **current demand** as closely as possible.

N. Herbst, S. Kounev and R. Reussner
Elasticity in Cloud Computing: What it is, and What it is Not.
in Proceedings of the 10th International Conference on Autonomic Computing (ICAC 2013), San Jose, CA, June 24-28, 2013.
[slides | http | .pdf]

http://en.wikipedia.org/wiki/Elasticity_(cloud_computing)

BUNGEE

- Problem: How to measure system elasticity (i.e., ability to automatically provision and deprovision resources to match the actual demand)?
- Framework for benchmarking elasticity
  - Current focus: IaaS cloud platforms

http://descartes.tools/bungee

Links for Further Information

- **DML** – Descartes Modeling Language (homepage, publications)
- **DML Bench** (homepage, publications)
- **DQL** – Declarative query language (homepage, publications)
- **DNI** – Descartes network infrastructure modeling (homepage, publications)
- **LibReDE** - Library for resource demand estimation (homepage, publications)
- **LIMBO** – Load intensity modeling tool (homepage, publications)
- **WCF** – Workload classification & forecasting tool (homepage, publications)
- **BUNGEE** – Elasticity benchmarking framework (homepage, publications)
- **hInjector** – Security benchmarking tool (homepage, publications)
- **Further relevant research**
  - **http://descartes-research.net/research/research_areas/**
  - **Self Aware Computing** (publications)

Coming Soon

Springer Book on
Self-Aware Computing Systems

Samuel Kounev (Universität Würzburg, DE)

Jeffrey O. Kephart (IBM TJ Watson, US)

Marta Kwiatkowska (University of Oxford, GB)

Xiaoyun Zhu (VMWare, Inc. – Palo Alto, US)

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Descartes
research

Questions?

skounev@acm.org
http://descartes.tools