Online Model-Based Performance and Resource Management in Virtualized Application Environments

Samuel Kounev

VMware Academic Research Symposium, Palo Alto, July 17, 2013

DESCARTES RESEARCH GROUP
INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION

www.kit.edu

References

Descartes
research

- Elastic Capacity Management / Online Workload Forecasting

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

  - N. Herbst, N. Huber, S. Kounev, and E. Amrehn. Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE 2013), Prague, Czech Republic, April 21-24*. 2013. [bib | slides | .pdf]

  - N. Huber, F. Brosig, and S. Kounev. Model-based Self-Adaptive Resource Allocation in Virtualized Environments. In *6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2011)*, Waikiki, Honolulu, HI, USA. May 23-24, 2011. [bib | http | .pdf]

- Automatic Model Extraction based on Benchmarking or Online System Monitoring

  - Q. Noorshams, D. Bruhn, S. Kounev, and R. Reussner. Predictive Performance Modeling of Virtualized Storage Systems using Optimized Statistical Regression Techniques. In *Proc. of the 4th ACM/SPEC International Conference on Performance Engineering*, Prague, Czech Republic, ICPE '13, pages 283-294, New York, NY, USA. ACM. 2013. [bib | DOI | http | .pdf]

  - F. Brosig, N. Huber, and S. Kounev. Automated Extraction of Architecture-Level Performance Models of Distributed Component-Based Systems. In *26th IEEE/ACM International Conference On Automated Software Engineering (ASE 2011)*, Oread, Lawrence, Kansas. November 2011. [bib | .pdf]

  - S. Kounev, K. Bender, F. Brosig, N. Huber, and R. Okamoto. Automated Simulation-Based Capacity Planning for Enterprise Data Fabrics. In *4th International ICST Conference on Simulation Tools and Techniques (SIMUTools 2011)*, Barcelona, Spain. March 21-25, 2011. **Best Paper Award**. [bib | slides | .pdf]

  - Q. Noorshams, K. Rostami, S. Kounev, P. Tůma, and R. Reussner. I/O Performance Modeling of Virtualized Storage Systems. In *Proceedings of the IEEE 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, San Francisco, USA. 2013. [bib | .pdf]

- Performance Modeling and Prediction in Virtualized Environments

  - F. Brosig, F. Gorsler, N. Huber, and S. Kounev. Evaluating Approaches for Performance Prediction in Virtualized Environments. In *Proceedings of the IEEE 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, San Francisco, USA. 2013. [bib | .pdf]

  - N. Huber, M. von Quast, M. Hauck, and S. Kounev. Evaluating and Modeling Virtualization Performance Overhead for Cloud Environments. In *Proceedings of the International Conference on Cloud Computing and Services Science (CLOSER 2011)*, Noordwijkerhout, The Netherlands, pages 563 - 573. SciTePress. May 7-9, 2011. **Best Paper Award**. [bib | http | .pdf]

References

- Descartes Meta-Model (DMM) - Online Models for Architecting Self-Aware Systems

  - http://www.descartes-research.net/research_and_profile/descartes_meta_model/

  - F. Brosig, N. Huber, and S. Kounev. Architecture-Level Software Performance Abstractions for Online Performance Prediction. *Elsevier Science of Computer Programming Journal (SciCo)*, 2013. [bib | DOI | .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 (SOCA), 2013, Springer. In print. [bib | .pdf]

  - N. Huber, A. van Hoorn, A. Koziolek, F. Brosig, and S. Kounev. S/T/A: Meta-Modeling Run-Time Adaptation in Component-Based System Architectures. In *9th IEEE Intl. Conf. on e-Business Engineering (ICEBE 2012), Hangzhou, China*, September 9-11, 2012. [bib | http | .pdf]

  - F. Brosig, N. Huber, and S. Kounev. Modeling Parameter and Context Dependencies in Online Architecture-Level Performance Models. In Proc. of the 15th ACM SIGSOFT Intl. Symposium on Component Based Software Engineering (CBSE 2012), June 26-28, 2012. [bib | .pdf]

  - N. Huber, F. Brosig, and S. Kounev. Modeling Dynamic Virtualized Resource Landscapes. In Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012), Bertinoro, Italy, June 25-28, 2012. [bib| .pdf]

  - S. Kounev, F. Brosig, and N. Huber. Descartes Meta-Model (DMM). Technical report, Karlsruhe Institute of Technology (KIT), 2013. To appear. [bib | http]

- Vision of Self-Aware Computing Systems

  - "Model-driven Algorithms and Architectures for Self-Aware Computing Systems" Dagstuhl Seminar scheduled to take place in October 2014 organized by Samuel Kounev, Jeff Kephart, Marta Kwiatkowska and Xiaoyun Zhu.

  - S. Kounev. Engineering of Self-Aware IT Systems and Services: State-of-the-Art and Research Challenges. In *Proc. of the 8th European Performance Engineering Workshop (EPEW'11), Borrowdale, The English Lake District, October 12-13*. 2011. (Keynote Talk). [bib | .pdf]

  - S. Kounev. Self-Aware Software and Systems Engineering: A Vision and Research Roadmap. In *GI Softwaretechnik-Trends, 31(4), November 2011, ISSN 0720-8928*, Karlsruhe, Germany, 2011. [bib | .html | .pdf]

  - S. Kounev, F. Brosig, and N. Huber. Towards self-aware performance and resource management in modern service-oriented systems. In *Proc. of the 7th IEEE Intl. Conference on Services Computing (SCC 2010), July 5-10, Miami, Florida, USA*. IEEE, 2010. [bib | .pdf]

References

- Cloud Usage Scenarios, Challenges and Opportunities

  - A. Milenkoski, A. Iosup, S. Kounev, K. Sachs, P. Rygielski, J. Ding, W. Cirne, and F. Rosenberg. Cloud Usage Patterns: A Formalism for Description of Cloud Usage Scenarios. Technical Report SPEC-RG-2013-001 v.1.0.1, SPEC Research Group - Cloud Working Group, Standard Performance Evaluation Corporation (SPEC), April 2013. [bib | .pdf]

  - S. Kounev, P. Reinecke, F. Brosig, J. T. Bradley, K. Joshi, V. Babka, A. Stefanek, and S. Gilmore. Providing dependability and resilience in the cloud: Challenges and opportunities. In K. Wolter, A. Avritzer, M. Vieira, and A. van Moorsel, editors, *Resilience Assessment and Evaluation of Computing Systems*, XVIII. Springer-Verlag, Berlin, Heidelberg, 2012. ISBN: 978-3-642-29031-2. [bib | http | .pdf]

- Performance Isolation in Shared Execution Environments (e.g., Multi-Tenant SaaS)

  - R. Krebs, C. Momm, and S. Kounev. Metrics and Techniques for Quantifying Performance Isolation in Cloud Environments. *Elsevier Science of Computer Programming Journal (SciCo)*, 2013. To appear. [bib]

  - R. Krebs, C. Momm, and S. Kounev. Metrics and Techniques for Quantifying Performance Isolation in Cloud Environments. In Barbora Buhnova and Antonio Vallecillo, editors, *Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012)*, Bertinoro, Italy, pages 91-100, New York, USA. ACM Press. June 25-28, 2012. [bib | http | .pdf]

- Intrusion Detection and Prevention in Virtualized Environments

  - A. Milenkoski, S. Kounev, A. Avritzer, N. Antunes, and M. Vieira. On Benchmarking Intrusion Detection Systems in Virtualized Environments. Technical Report SPEC-RG-2013-002 v.1.0, SPEC Research Group - IDS Benchmarking Working Group, Standard Performance Evaluation Corporation (SPEC), June 2013. [bib | .pdf]
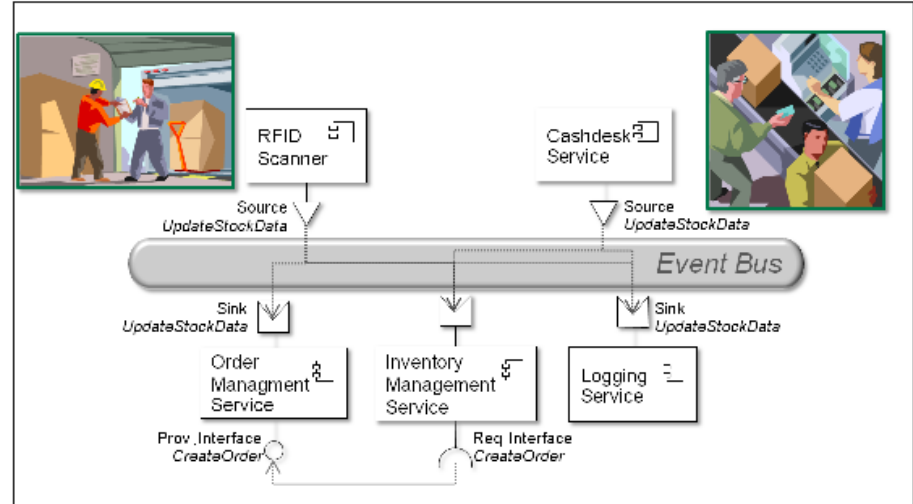
Agenda

- Motivation

- Goals & Approach

- Case Studies

  - Workload Forecasting

  - Performance Modeling

  - Resource Management

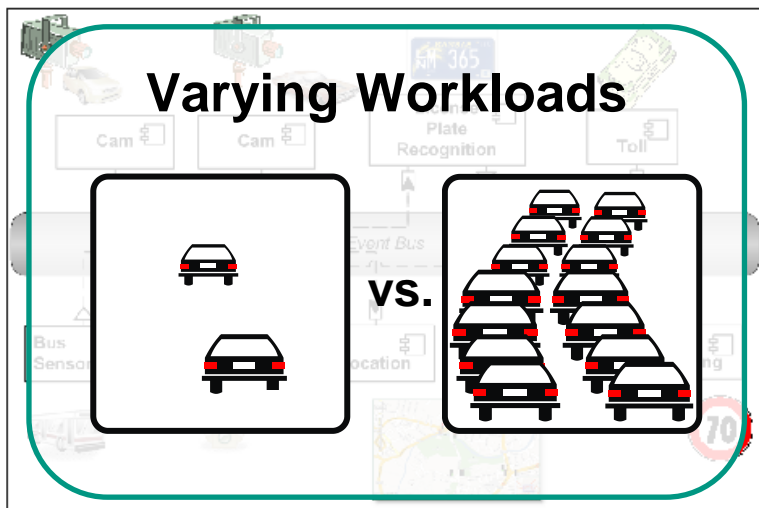- Outlook

Motivation

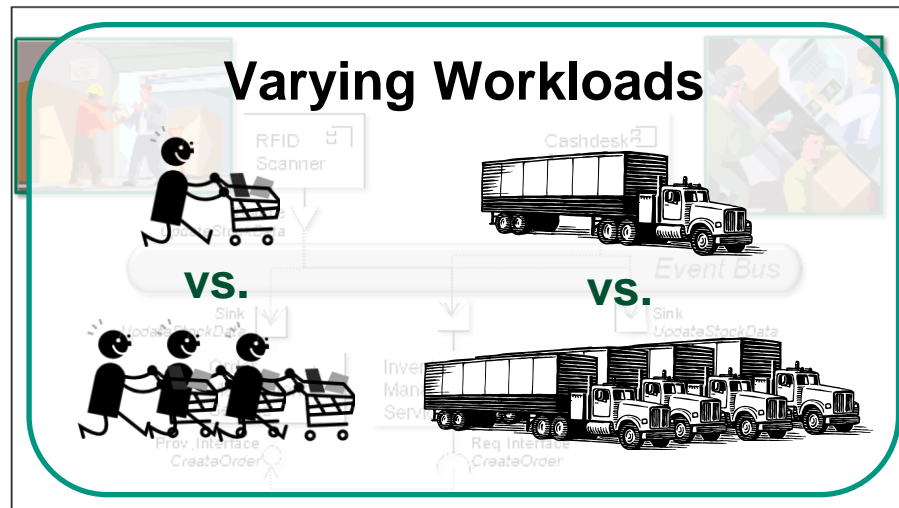## Traffic Monitoring System
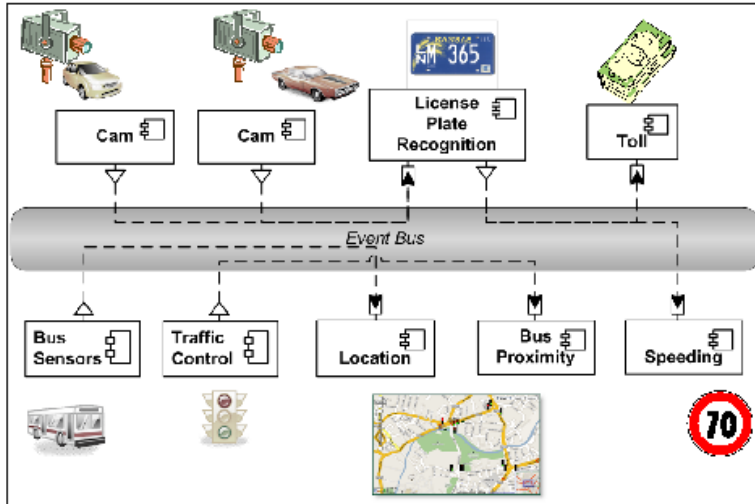
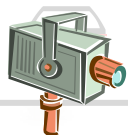## Inventory Management System

Motivation

Traffic Monitoring System

Inventory Management System

Motivation

Traffic Monitoring System

System Evolution
- New streets / bus lines
- New features and services
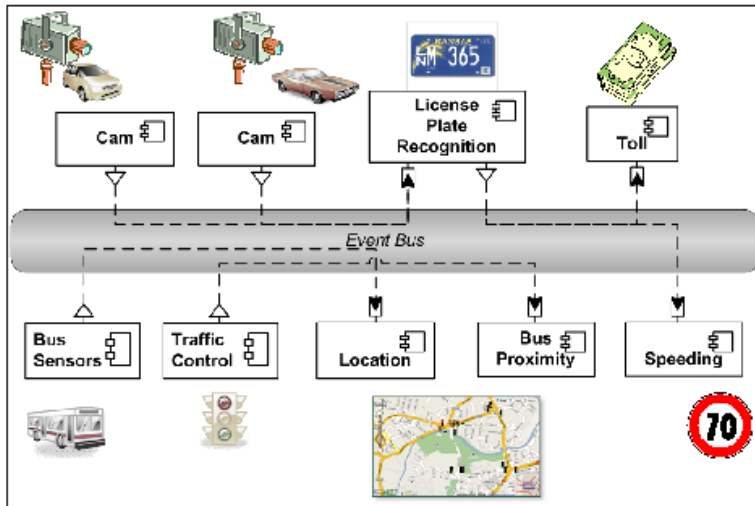- Upgraded cameras

vs.

Inventory Management System

System Evolution
- New supermarket stores
- New features and services
- Upgraded RFID readers

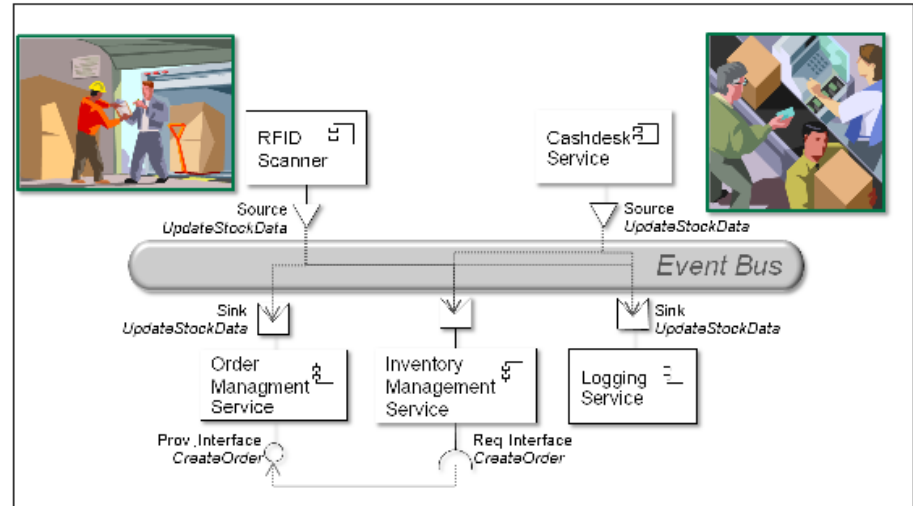**vs.**

Motivation

Traffic Monitoring System

Inventory Management System

- Software systems increasingly **complex** and **dynamic**

- Must be **reconfigured at run-time** more and more frequently

  - Component instances, application configuration

  - Deployment topology, resource allocations

- Two issues:

  - Determine **WHEN** exactly reconfigurations are necessary?

  - Determine **WHAT** exactly each reconfiguration should do?

Example: Elastic Resource Provisioning

SLAs

Load Fluctuations

Application 1

Application 2

Application N

VAN

VMM

VAN

VMM

VAN

VMM

Example: Elastic Resource Provisioning

Load Fluctuations
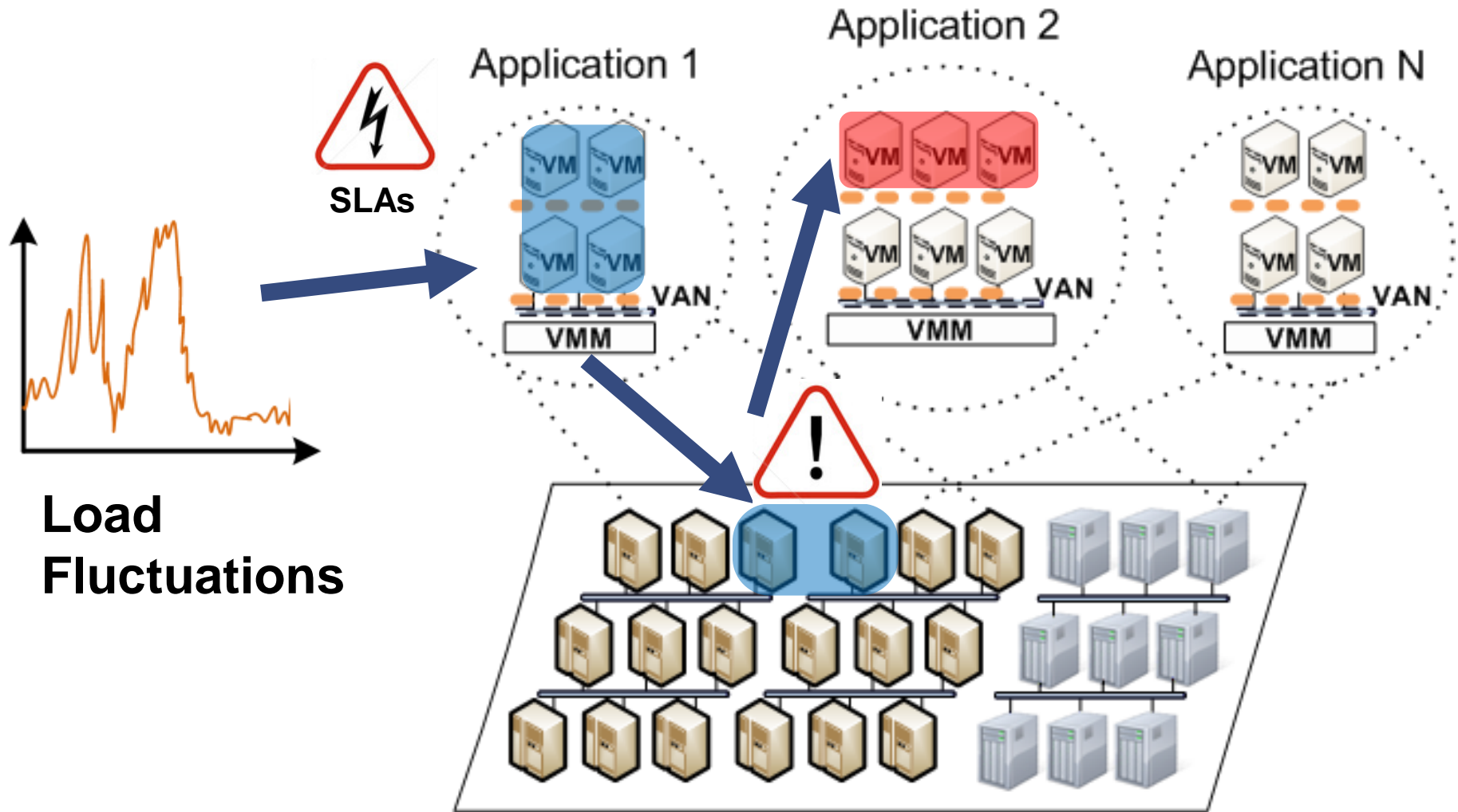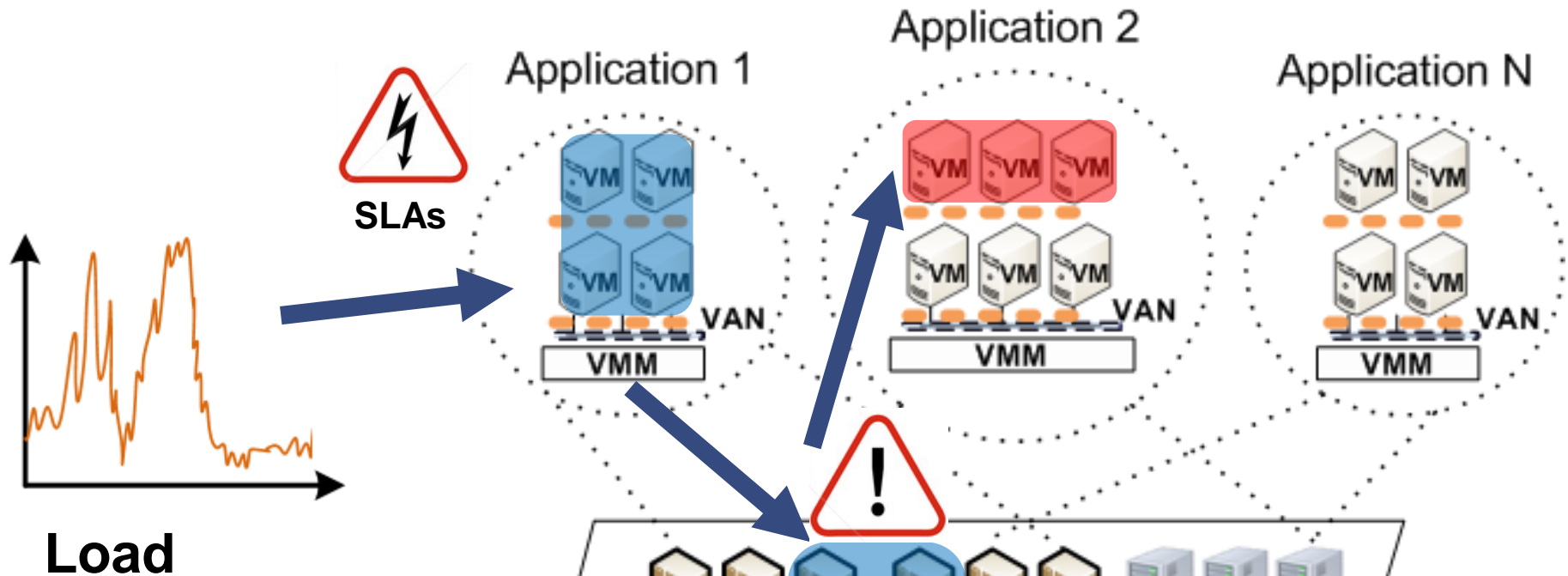
Expand / shrink resources on-the-fly

- When exactly should a reconfiguration be triggered?

- Which particular resources should be scaled?

- How quickly and at what granularity?

Semantic Gap Problem

Applications 🟧 🟥 🟩
- Multiple tiers
- Multiple resource types

| VM 🟧 | VM 🟩 | VM 🟩 |

VMM

Host

Resource Allocation

| VM 🟧 | VM 🟥 | VM 🟩 |

VMM

Host

| VM | VM | VM |

VMM

Host

| EAR | EAR |
| Java EE |
| JVM |
| OS |
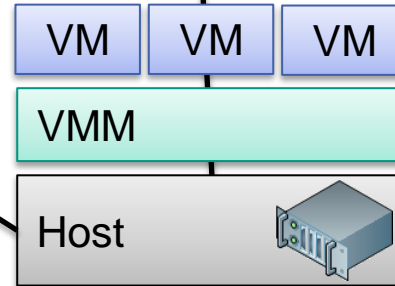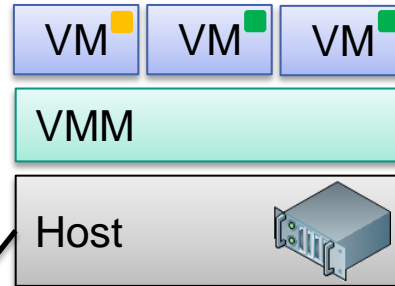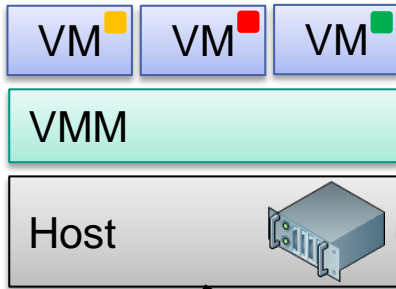
Complex Software Stacks
- Multiple layers
- Heterogeneous

| End-to-end QoS metrics Application SLAs | **?** | Resource Allocations in Each Tier & Each Layer |

State-of-the-Art

- Hard to predict the effect of dynamic changes on the system performance and resource demands

    - Rely on simple trigger-based "best effort" adaptation mechanisms OR
    - Avoid need for adaptation by over-provisioning resources

- Consequences: Poor resource efficiency

    - Rising energy costs for IT systems

        - 1600% increase by 2025 [Gartner]

    - Rising global CO_2 emissions of ICT sector

        - Today: ca 3%, Increase to 10% expected in 10 years [EU]

Descartes Research Group

- Modeling methods for **predicting at run-time** the effect of dynamic changes on the system Quality-of-Service (QoS)

  - Current focus: availability and performance (response time, throughput and resource/energy efficiency)

- Model-based algorithms and techniques for **autonomic system adaptation** during operation

- Goal:

  - End-to-end QoS guarantees

  - High resource/energy efficiency

  - Low operating costs

Descartes Research Group

Self-Aware Software Systems

- ## Self-Reflective

  - Aware of their software architecture, execution environment and hardware infrastructure, as well as of their operational goals (e.g., QoS and efficiency)

- ## Self-Predictive

  - Able to anticipate and predict the effect of dynamic changes in the environment, as well as the effect of possible adaptation actions

- ## Self-Adaptive

  - Proactively adapting as the environment evolves to ensure that their operational goals are continuously met

"I think, therefore I am…"
-- René Descartes

Self-Aware Software Systems

Self-Reflective

- Aware of their software architecture, execution environment and hardware infrastructure, as well as of their operational goals (e.g., QoS and efficiency)

Self-Predictive

- Able to anticipate and predict the effect of dynamic changes in the environment, as well as the effect of possible adaptation actions

Self-Adaptive

- Proactively adapting as the environment evolves to ensure that their operational goals are continuously met

"Model-driven Algorithms and Architectures for Self-Aware Computing Systems" Dagstuhl Seminar scheduled to take place in October 2014 organized by Samuel Kounev, Jeff Kephart, Marta Kwiatkowska and Xiaoyun Zhu.

S. Kounev. Engineering of Self-Aware IT Systems and Services: State-of-the-Art and Research Challenges. In *Proc. of the 8th European Performance Engineering Workshop (EPEW'11), Borrowdale, The English Lake District, October 12-13.* 2011. (Keynote Talk). [bib | .pdf]

S. Kounev. Self-Aware Software and Systems Engineering: A Vision and Research Roadmap. In *GI Softwaretechnik-Trends, 31(4), November 2011, ISSN 0720-8928*, Karlsruhe, Germany, 2011. [bib | .html | .pdf]

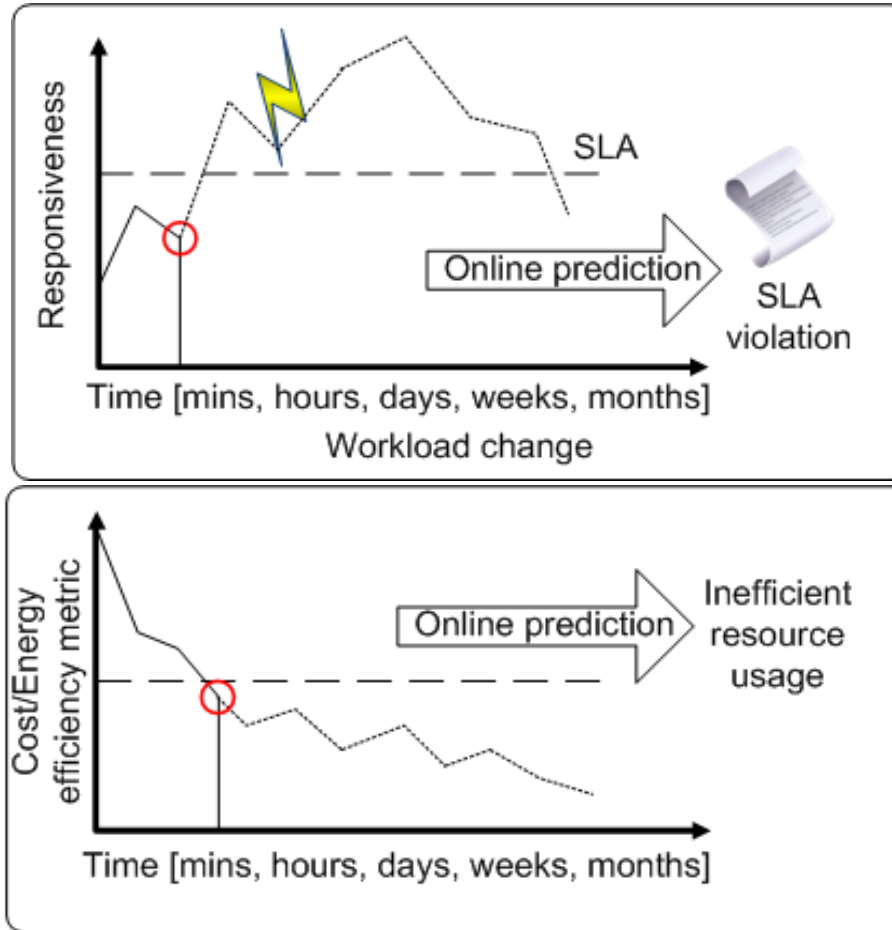S. Kounev, F. Brosig, and N. Huber. Towards self-aware performance and resource management in modern service-oriented systems. In *Proc. of the 7th IEEE Intl. Conference on Services Computing (SCC 2010), July 5-10, Miami, Florida, USA*. IEEE, 2010. [bib | .pdf]

Vision: Self-Aware Computing Systems
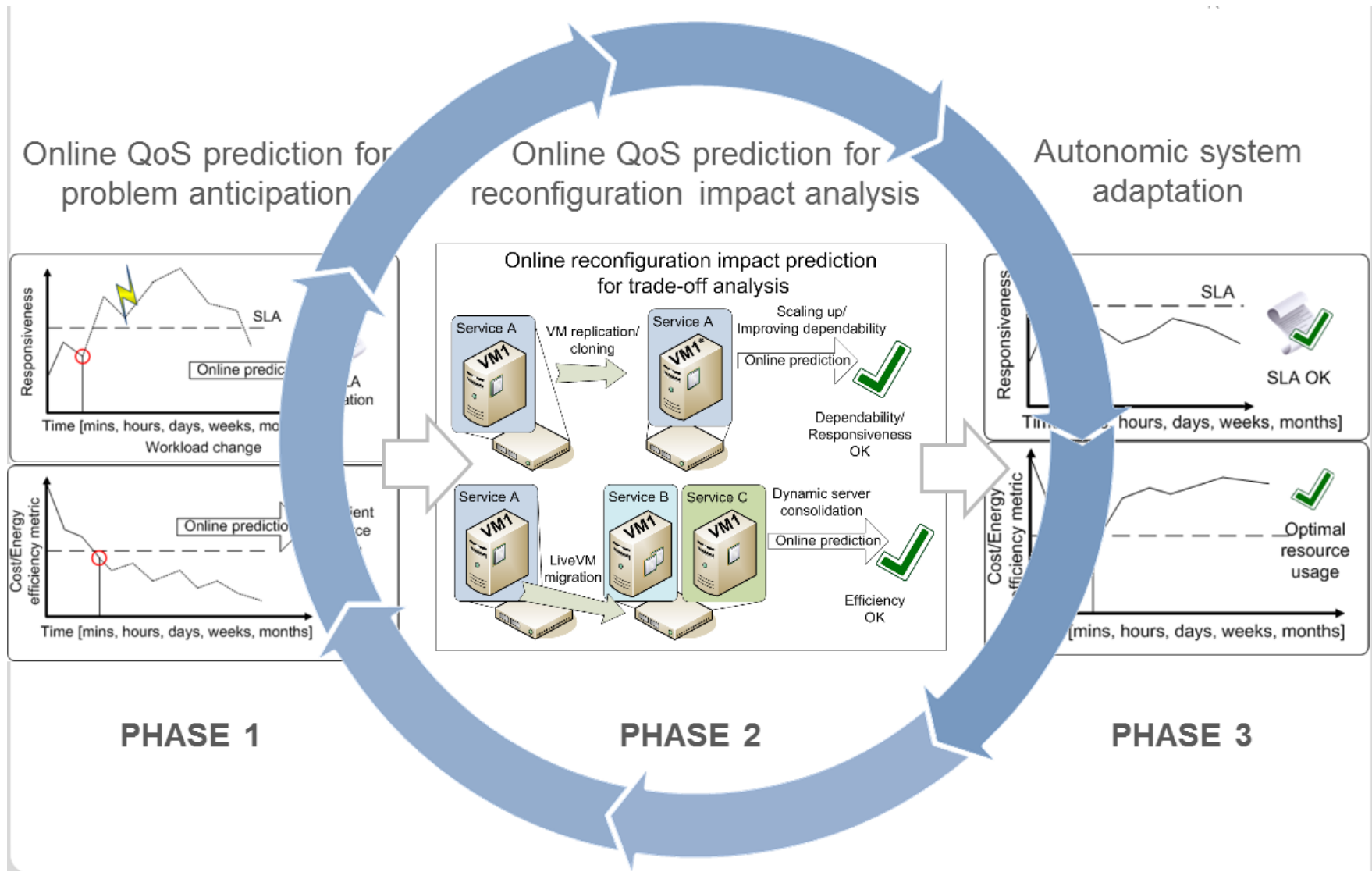
Vision: Self-Aware Computing Systems

PHASE 1
Online QoS Prediction for Problem Anticipation

Vision: Self-Aware Computing Systems

Online reconfiguration impact prediction for trade-off analysis

Service A — VM1 — VM replication/ cloning → Service A — VM1* — Online prediction ✓

Scaling up/ Improving dependability

Dependability/ Responsiveness OK

Service A — VM1 — LiveVM migration → Service B — VM1, Service C — VM1 — Online prediction ✓

Dynamic server consolidation

Efficiency OK
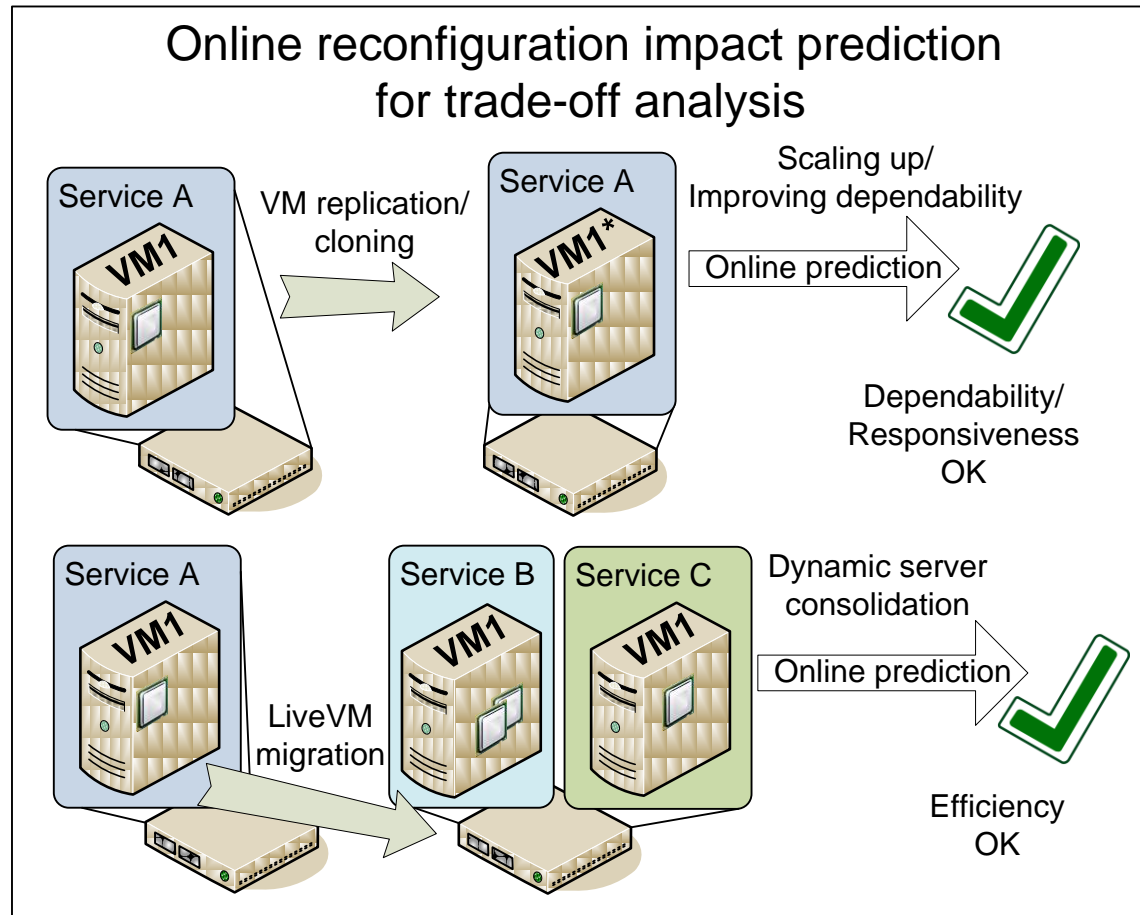
PHASE 2
Online QoS Prediction for Reconfiguration Impact Analysis

Motivation ⟩⟩ **GOALS & APPROACH** ⟩⟩ Case Studies ⟩⟩ Outlook

© S. Kounev   Online Model-Based Performance and Resource Management in Virtualized Application Environments

Vision: Self-Aware Computing Systems
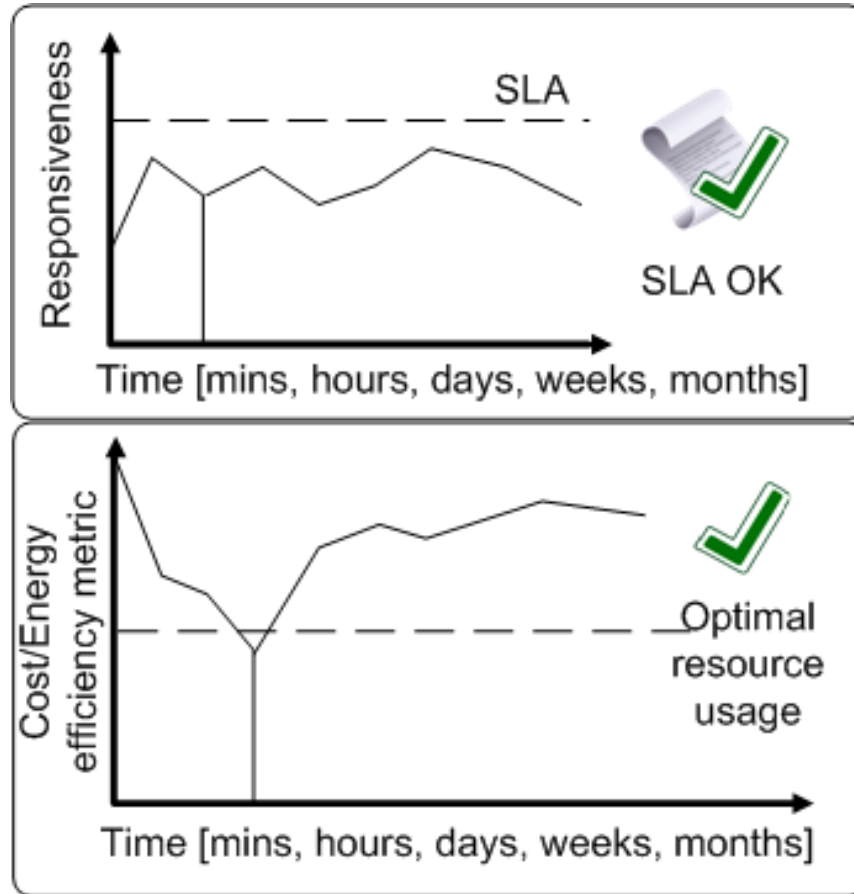
PHASE 3
Autonomic System Adaptation

Vision: Self-Aware Computing Systems

Examples of Performance-Influencing Factors

System workload and usage profile
- Number and type of clients
- Input parameters and input data
- Data formats used
- Service workflow

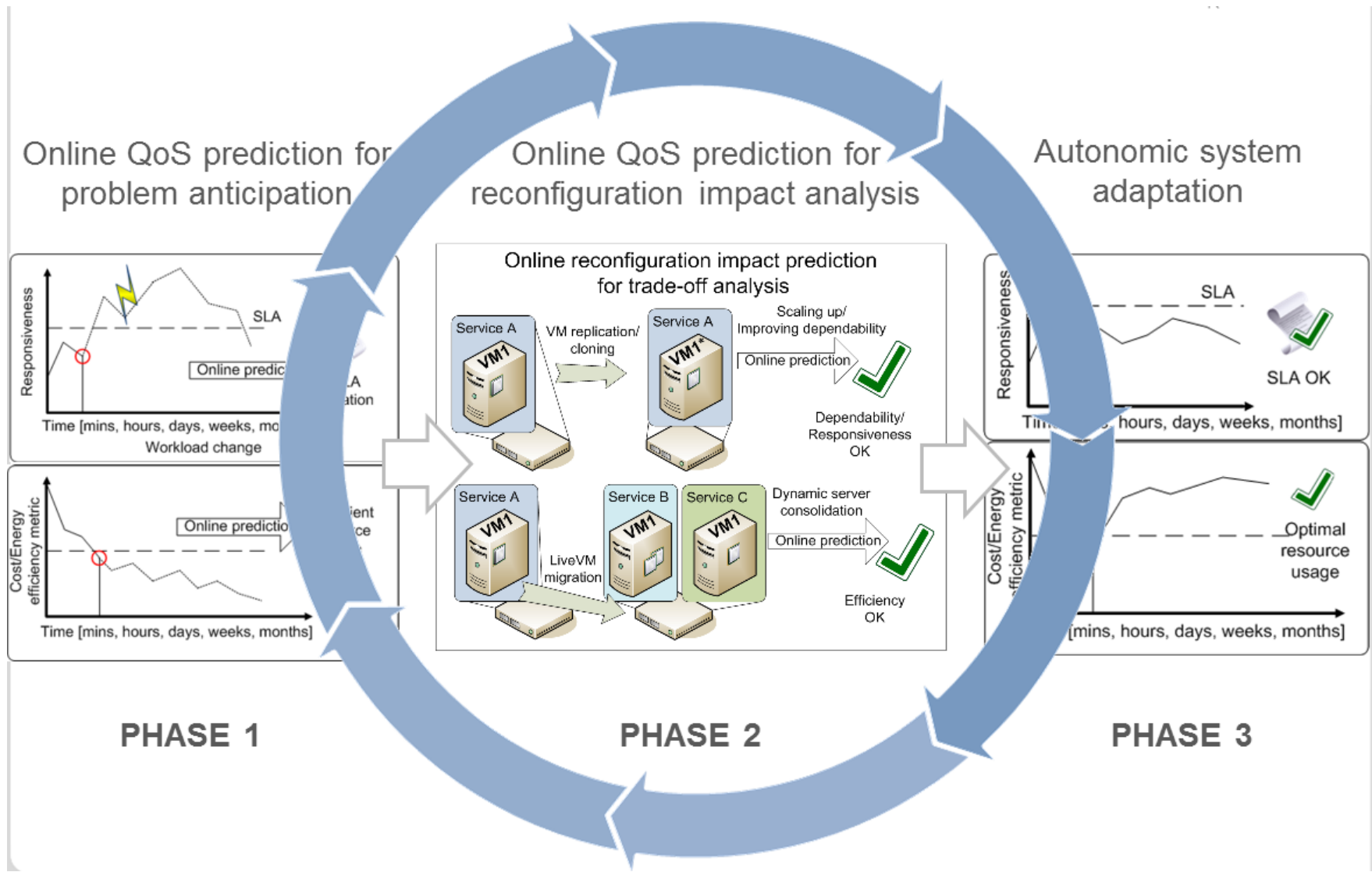**Software architecture**
- Connections between components
- Flow of control and data
- Component resource demands
- Component usage profiles

Execution environment
- Number of component instances
- Server execution threads
- Amount of Java heap memory
- Size of database connection pools

Virtualization layer
- Physical resources allocated to VMs
  - number of physical CPUs
  - amount of physical memory
  - secondary storage devices
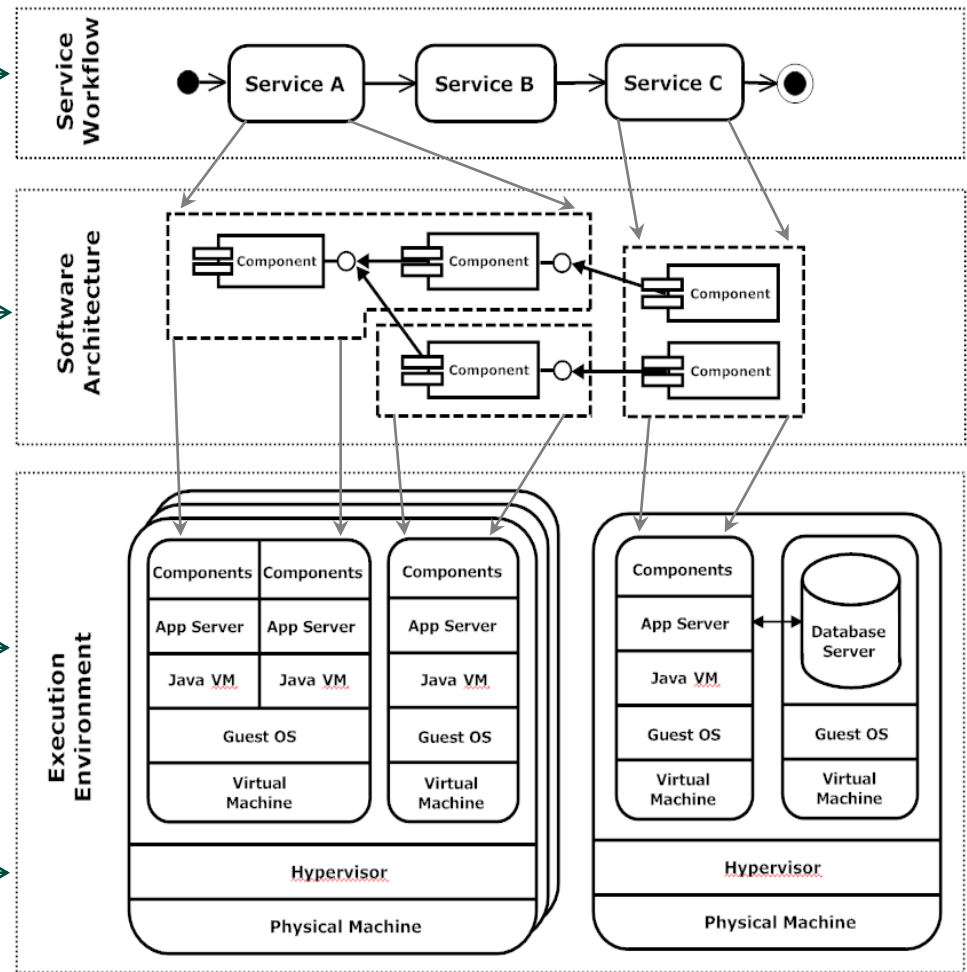
Network bandwidth between system nodes

Descartes Meta-Model (DMM)

- Architecture-level modeling language for modeling QoS and resource management related aspects of IT systems
  - Prediction of the impact of dynamic changes at run-time
  - Current version focused on performance (incl. capacity and efficiency)
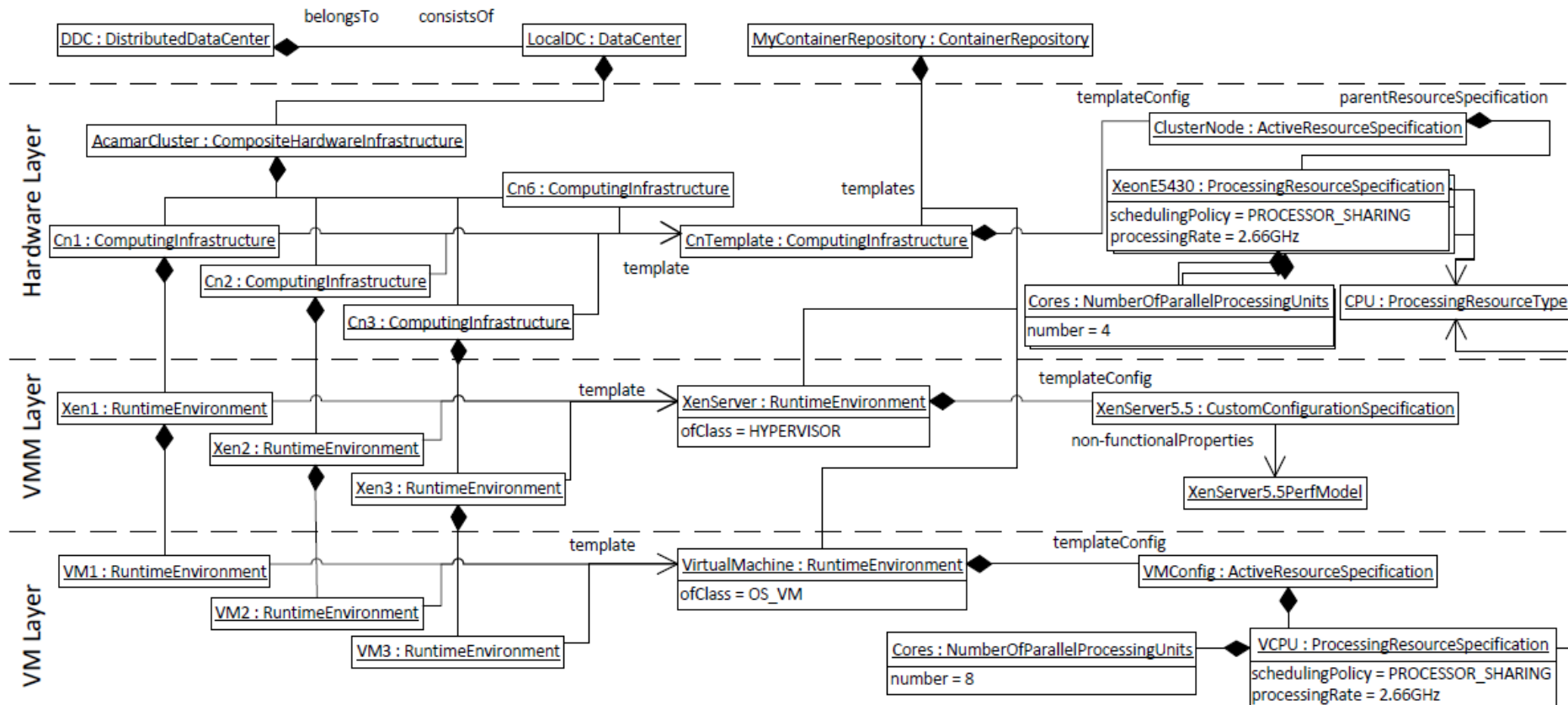
[Springer SOCA 2013]
[IEEE ICEBE 2012]

[Elsevier SciCo 2013]
[ACM CBSE 2012]
[IEEE/ACM ASE 2011]

[ACM QoSA 2012]
[CLOSER 2011]
[DOA 2010]

[ACM QoSA 2012]

Technical Report
+
IEEE TSE Article Under Review

Example: Resource Environment

N. Huber, F. Brosig, and S. Kounev. **Modeling Dynamic Virtualized Resource Landscapes**. In Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012), Bertinoro, Italy, June 25-28, 2012. [bib| .pdf]

Motivation　≫　**GOALS & APPROACH**　≫　Case Studies　≫　Outlook

Online Model-Based Performance and Resource Management in Virtualized Application Environments

Example: Resource Environment
Influence Factors of the Virtualization Layer

Descartes *research*

Legend
- ◇ exclusive OR
- ◆ inclusive OR

Virtualization Platform

Virtualization Type
- Binary Translation
- Full Virtualization
- Para-Virtualization

VMM Architecture
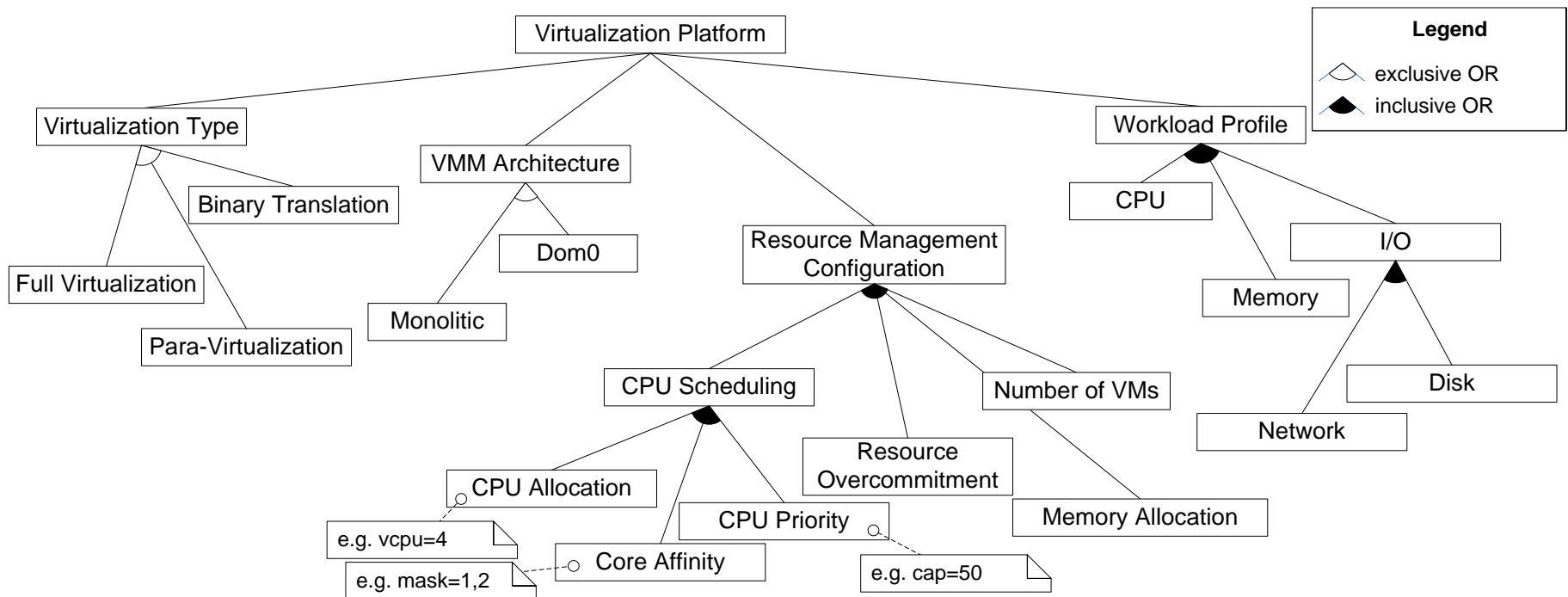- Dom0
- Monolitic

Resource Management Configuration
- CPU Scheduling
  - CPU Allocation
    - e.g. vcpu=4
    - e.g. mask=1,2
  - Core Affinity
  - CPU Priority
    - e.g. cap=50
- Resource Overcommitment
  - Memory Allocation
- Number of VMs

Workload Profile
- CPU
- Memory
- I/O
  - Network
  - Disk

N. Huber, M. von Quast, M. Hauck, and S. Kounev. **Evaluating and Modeling Virtualization Performance Overhead for Cloud Environments**. In *Proceedings of the International Conference on Cloud Computing and Services Science (CLOSER 2011)*, Noordwijkerhout, The Netherlands, pages 563 - 573. SciTePress. May 7-9, 2011. **Best Paper Award**. [bib | http | .pdf]

Motivation 〉〉 **GOALS & APPROACH** 〉〉 Case Studies 〉〉 Outlook

 Online Model-Based Performance and Resource Management in Virtualized Application Environments

Example: Application Architecture

- Control flow and data flow, service resource demands
- Parameter and context dependencies
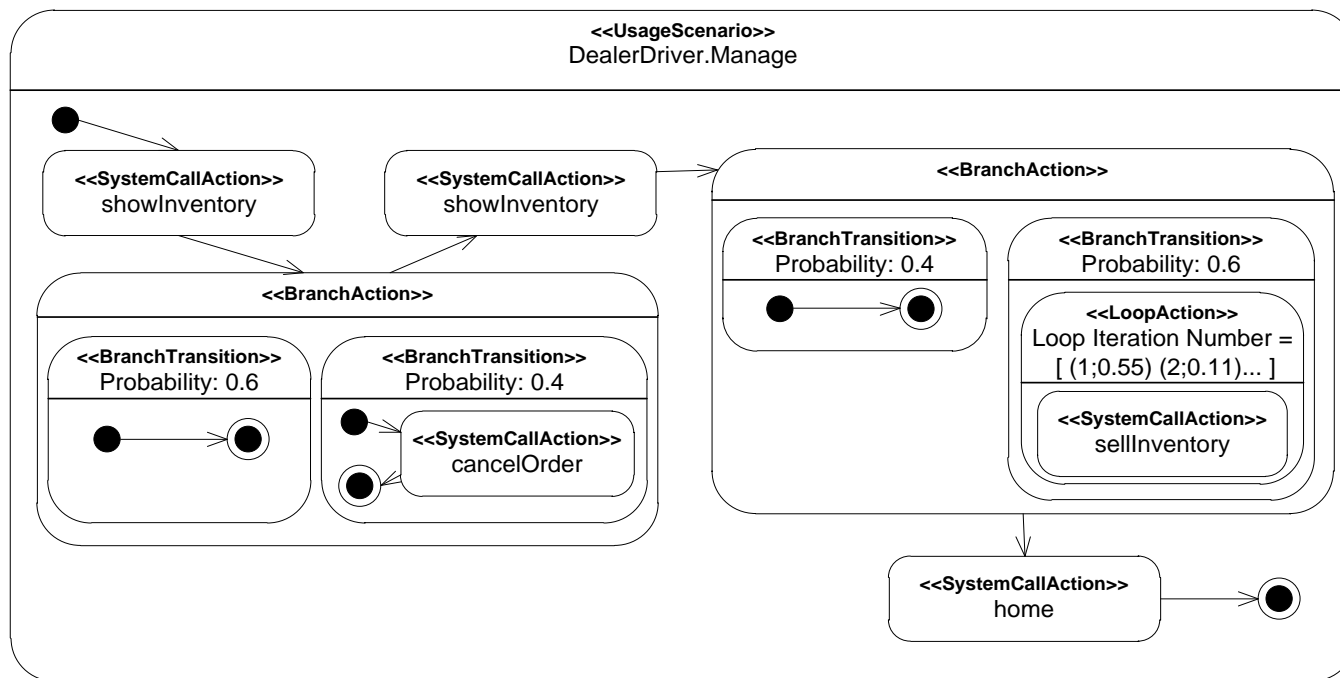
F. Brosig, N. Huber, and S. Kounev. **Architecture-Level Software Performance Abstractions for Online Performance Prediction**. *Elsevier Science of Computer Programming Journal (SciCo)*, 2013. [bib | DOI | http | .pdf]

F. Brosig, N. Huber, and S. Kounev. **Modeling Parameter and Context Dependencies in Online Architecture-Level Performance Models**. *15th ACM SIGSOFT Intl. Symposium on Component Based Software Engineering (CBSE 2012),* June 26-28, 2012. [bib | http | .pdf | Abstract]

Impact Prediction through Automatic Generation of Tailored Predictive Models at Run-Time

Online Model-Based Performance and Resource Management in Virtualized Application Environments
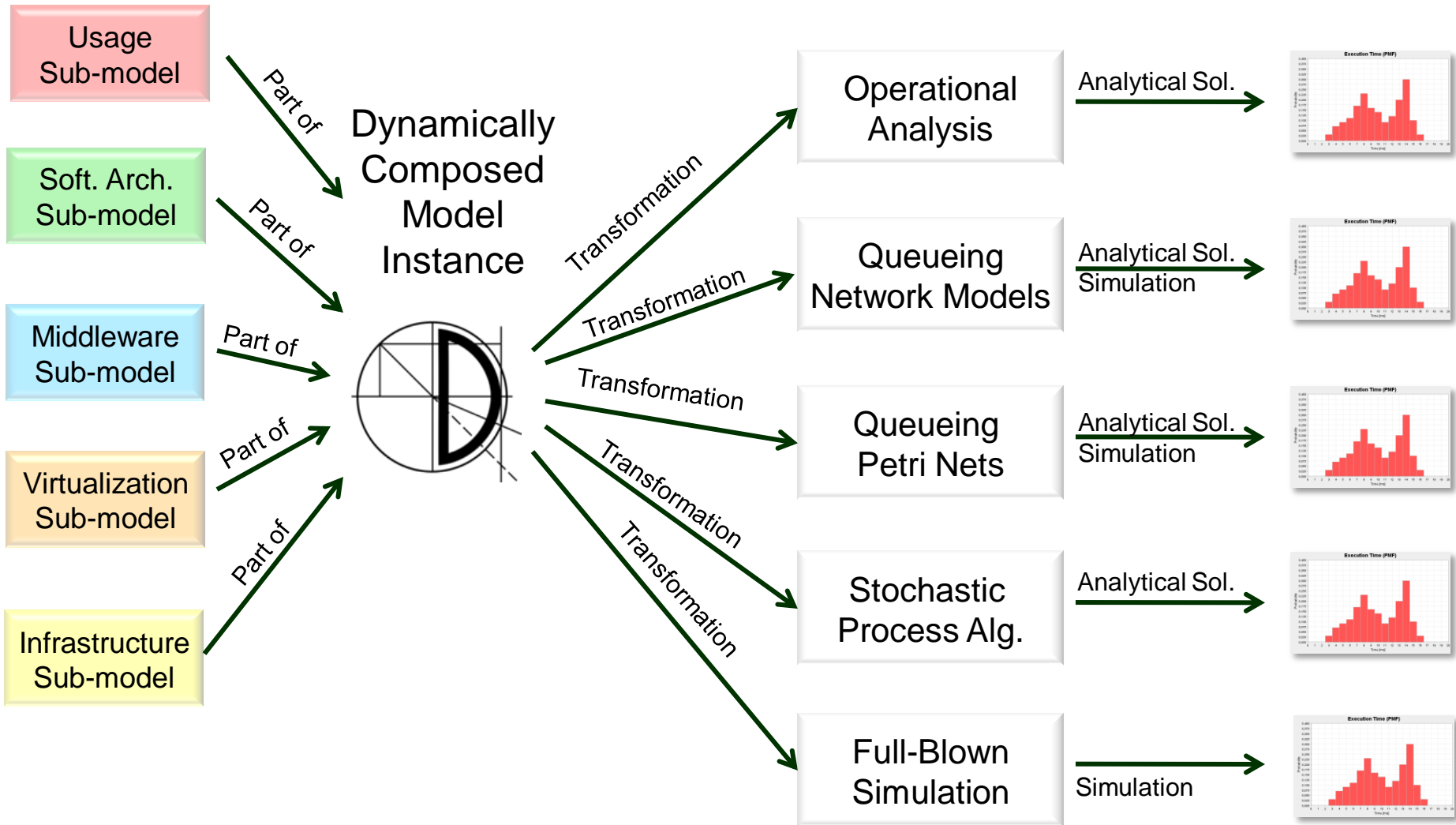
Case Study: Process Control System (ABB)
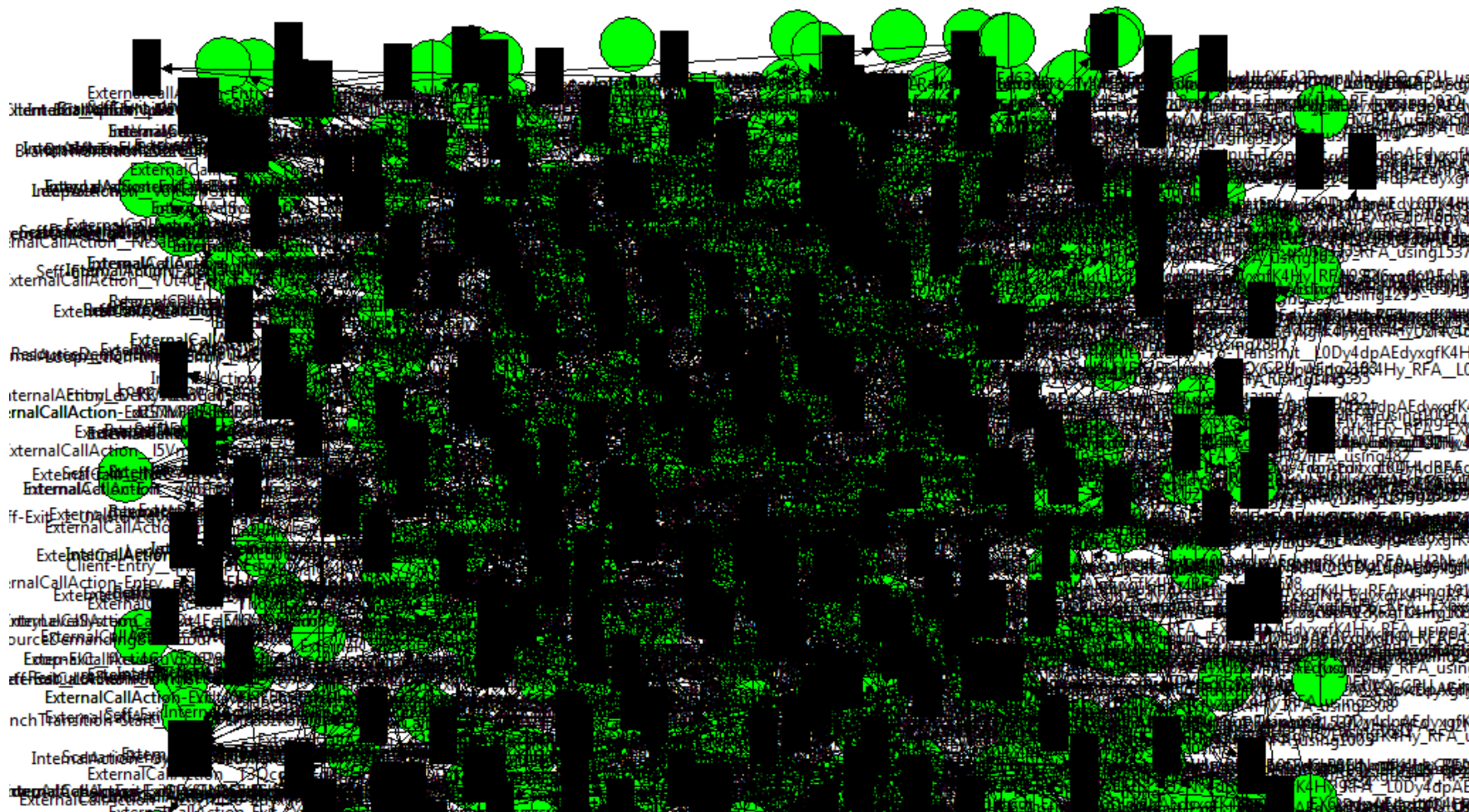
P. Meier, S. Kounev, and H. Koziolek. **Automated Transformation of Component-based Software Architecture Models to Queueing Petri Nets**. In *19th IEEE/ACM Intl. Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), Singapore, July* 2011. [bib | .pdf]

Overview of Applied Modeling Techniques

**Descriptive
Architecture-level Models**

- OMG Meta Object Facility (MOF)
  - MOF-based meta-models
- (UML MARTE)
- (UML SPT)

Predictive Performance Models

- Bounding techniques
- Operational analysis
- Statistical regression models
- Stochastic process algebras
- (Extended) queueing networks
- Layered queueing networks
- Queueing Petri nets
- Machine learning-based models
- Detailed simulation models

**Workload
Forecasting**

- AR(I)MA
- Extended exp. smoothing
- tBATS
- Croston's method
- Cubic smoothing splines
- Neural network-based

**Resource
Demand
Estimation**

- Regression-based techniques
- Kalman filter
- Nonlinear optimization
- Maximum likelihood estimation
- Independent component analysis

**Regression
Analysis**

- MARS
- CART
- M5 trees
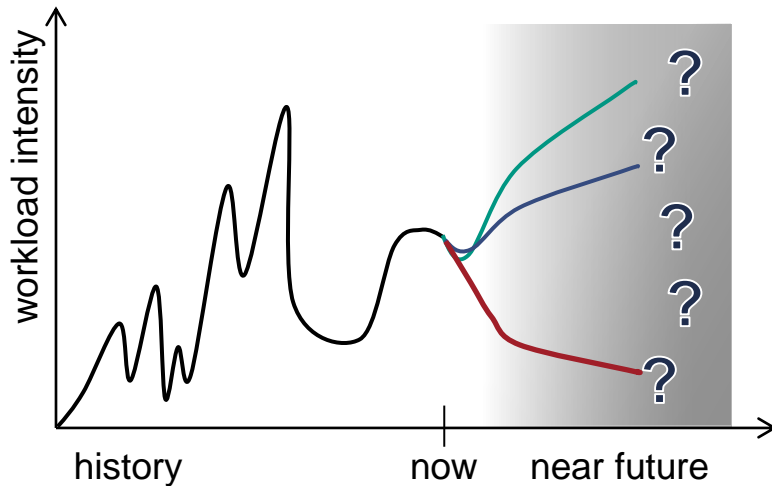- Cubist forests
- Quantile regression forests
- Support vector machines

Motivation ⟩⟩ **GOALS & APPROACH** ⟩⟩ Case Studies ⟩⟩ Outlook

© S. Kounev   Online Model-Based Performance and Resource Management in Virtualized Application Environments

Case Study 1: Workload Forecasting

- Use multiple alternative forecasting methods in parallel
- Select which method to trust based on its accuracy in recent time horizons

N. R. Herbst, N. Huber, S. Kounev, and E. Amrehn. **Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning**. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE 2013), Prague, Czech Republic, April 21-24.* 2013. [bib | slides | .pdf]

Time Series Analysis

[BFAST]
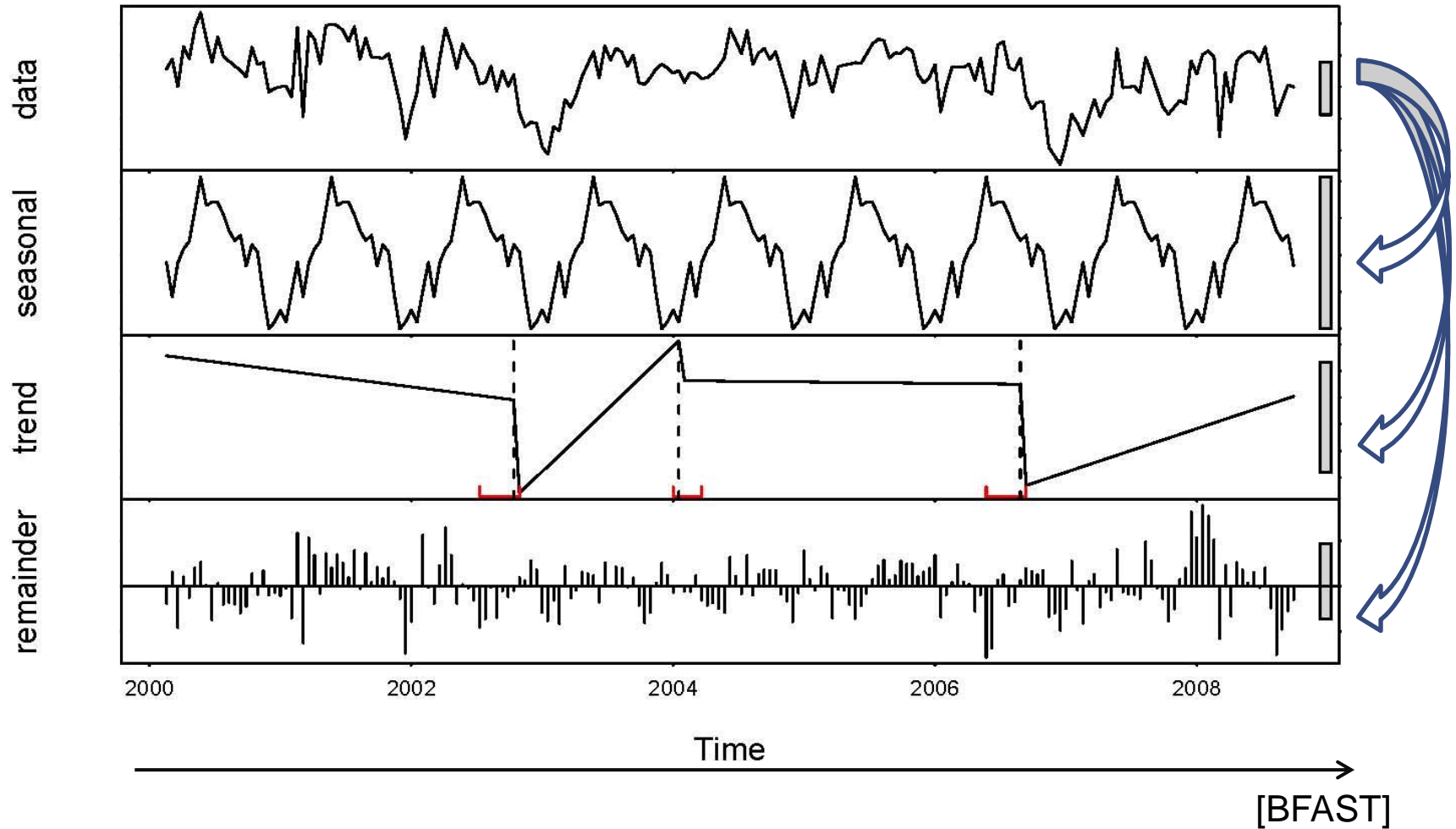
Forecasting Methods Used

Basic Methods	(initial)
Naïve, Moving Averages, Random Walk	

Trend Interpolation	(fast)
Simple Exponential Smoothing (SES)	[Hynd08]
Cubic Smoothing Splines	[Hynd02]
Croston's method for intermittent time series	[Shen05]
Autoregressive Moving Averages (ARMA11)	[Box08]

Estimation and Modelling of Seasonal Pattern	(complex)
Extended Exponential Smoothing (ETS)	[Hynd08, Hyn08]
ARIMA framework with automatic model selection	[Box08, Hynd08]
tBATS for complex seasonal patterns	[Live11]

WCF Technique
Self-Adaptive Workload Classification & Forecasting

Descartes
research

Overhead Constraint	Forecast Frequency
Forecast Objectives	
Forecast Horizon	Confidence Level

Forecast Objectives

Workload Intensity Trace

Accuracy Feedback

(II) Forecast & Evaluate Accuracy

(I) Select Method

Forecast Methods Overhead Groups

INITIAL Naïve, Smoothing, Random Walk

FAST Trend Inter-polation

COMPLEX Decom-position & Seasonal Patterns

Workload Intensity Characterization

High level data analysis to gain information on:

- Noise level & occurrences of unpredictable bursts
- Influence of trends and seasonal patterns
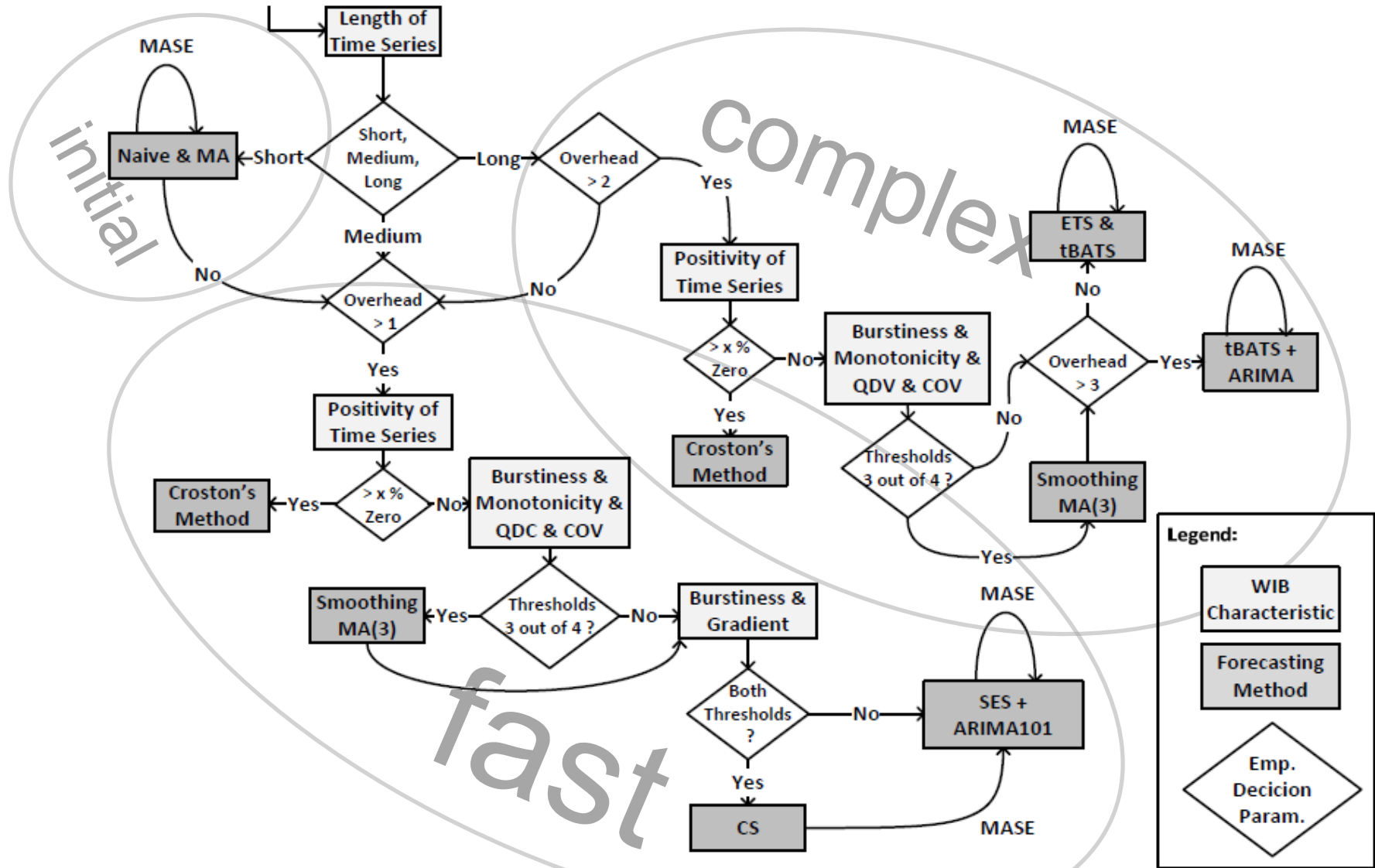

Data

| Workload Intensity Trace: Time Series of Request Arrival Rates |

Attributes: Length, Positivity, Frequency

Metrics: Gradient, Mono-tonicity, Variance, Quartile Dispersion, Burstiness
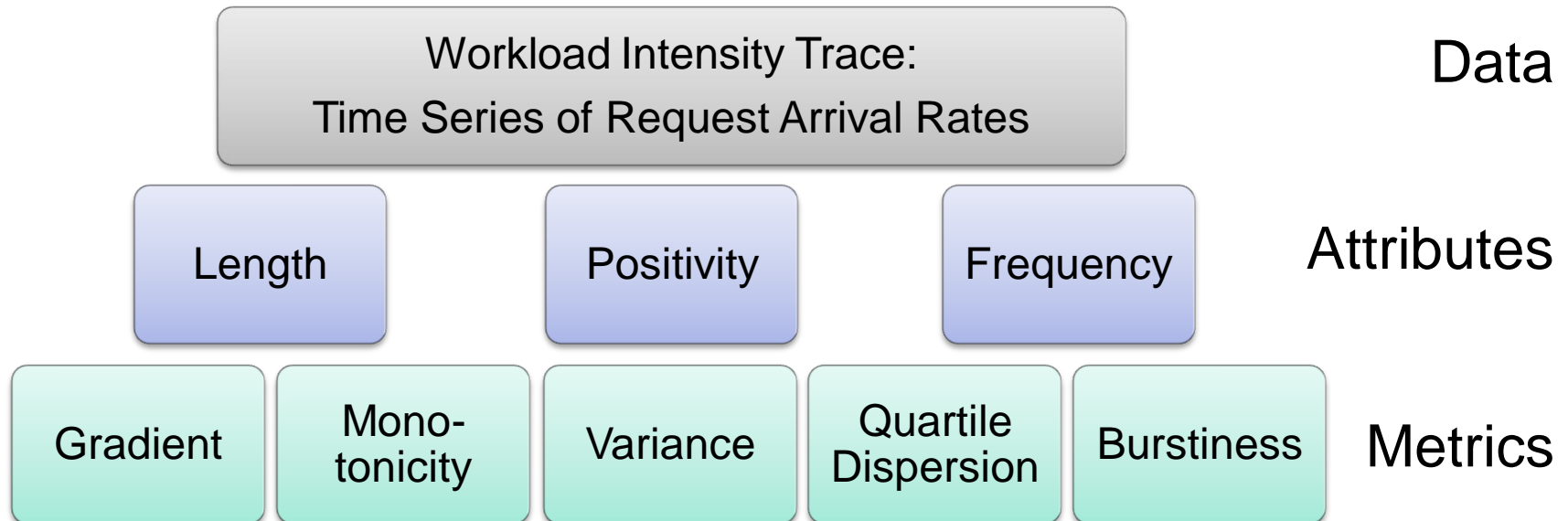
Decision Tree for Classification

Process of Classification & Forecasting

classification:
(I) select strategies

classification:
(II) evaluate selection

new classification:
(I) select strategies

forecast
execution &
result output

forecast
execution &
result output

forecast
execution &
result output

forecast
execution &
result output

…

forecast horizon/
frequency
(need not to be equal)

new forecast
horizon

classification frequency (# forecast executions)
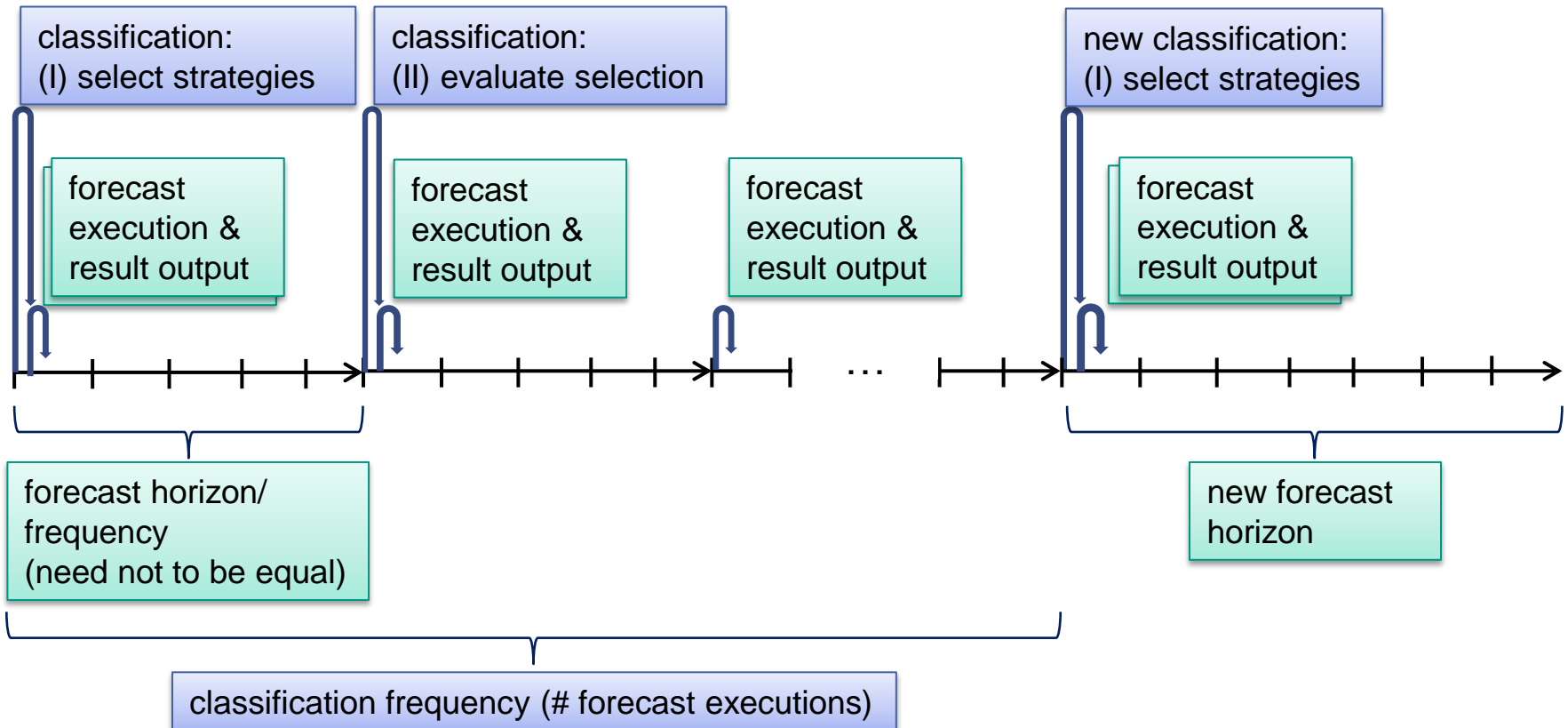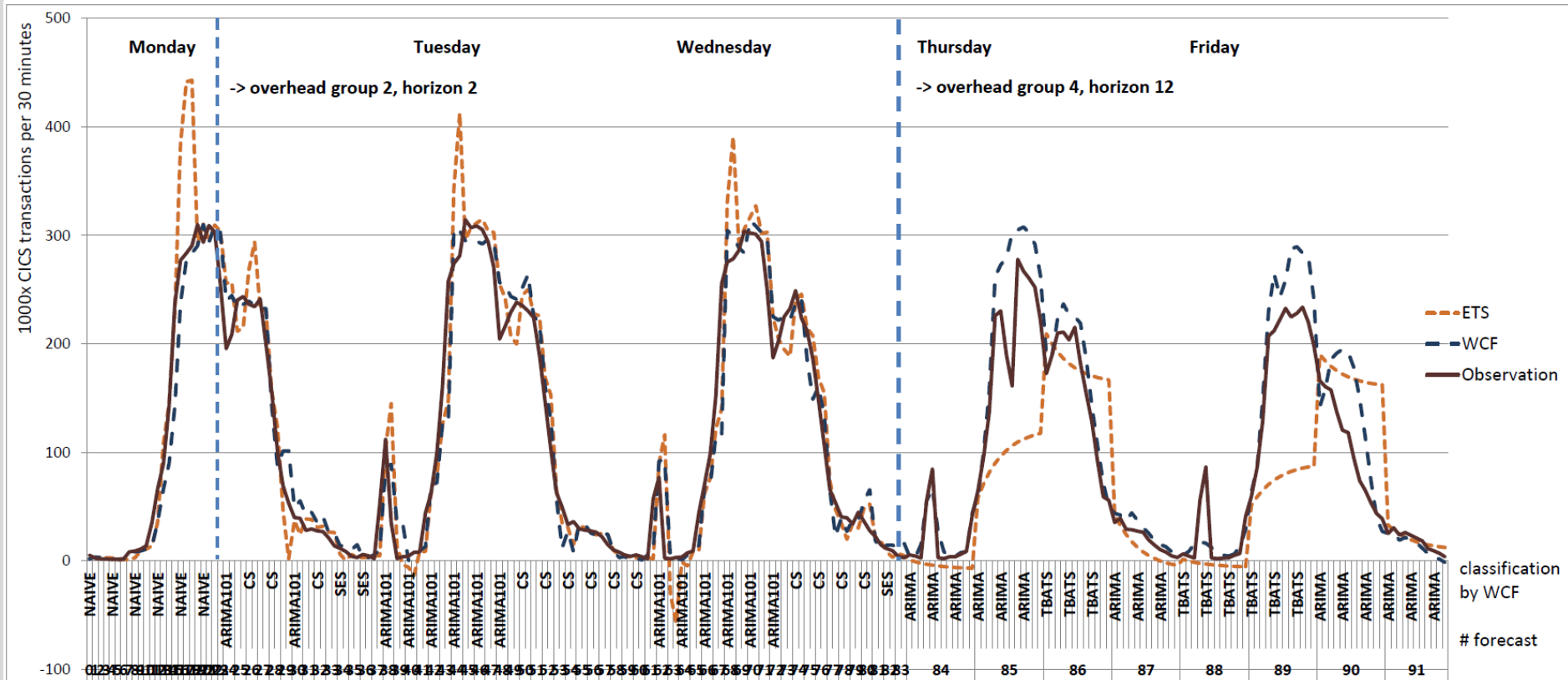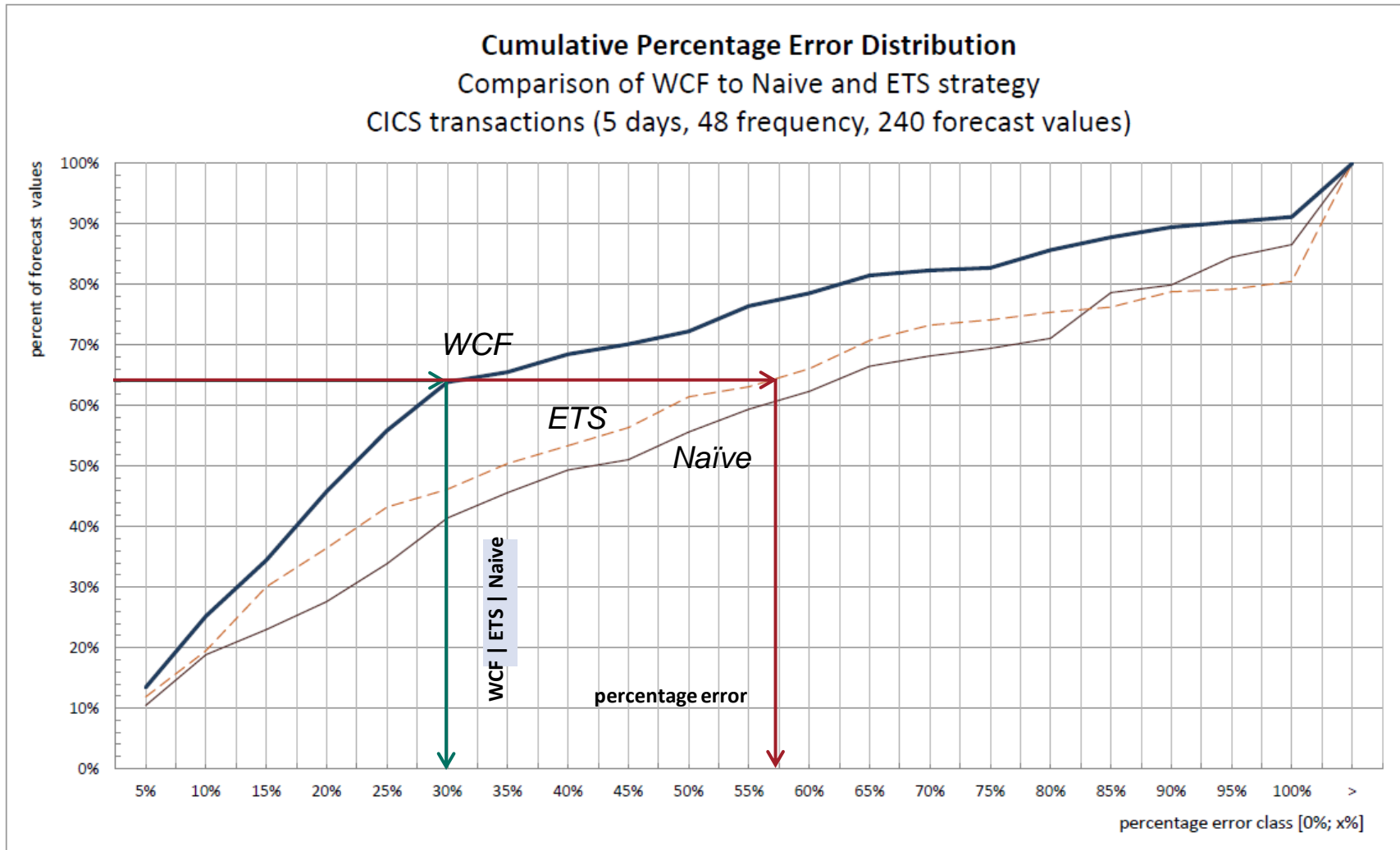
Forecast Accuracy Improvement

- Real-world workload intensity trace: IBM CICS transactions on System z
- Comparison of **WCF** approach to **ETS** forecast

Forecast Accuracy Improvement

Cumulative Percentage Error Distribution
Comparison of WCF to Naive and ETS strategy
CICS transactions (5 days, 48 frequency, 240 forecast values)
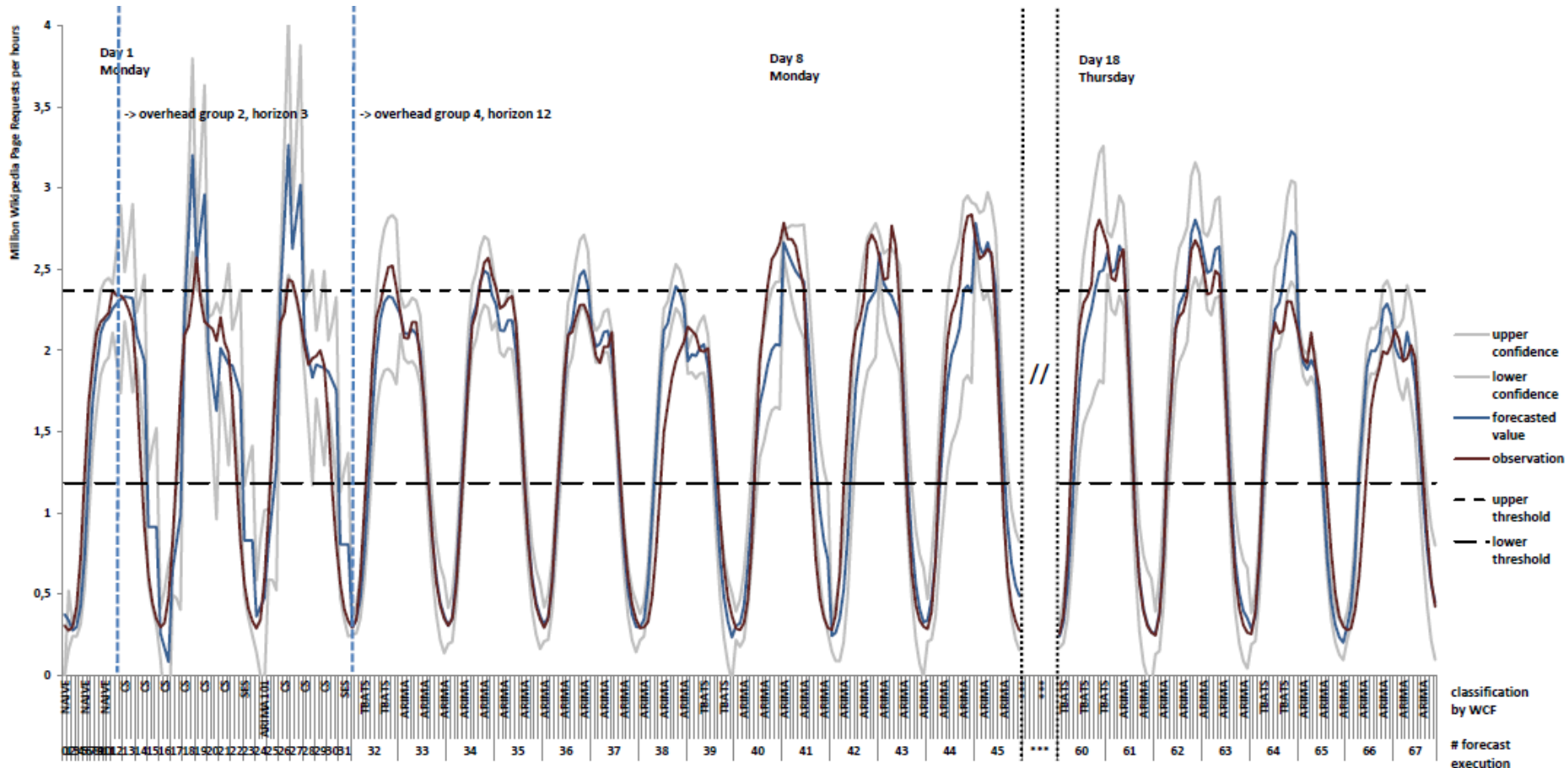
Case Study:
Example for Using Forecast Results

- **Scenario:** Additional server instances at certain thresholds, 3 weeks
- Real-world workload intensity trace (**Wikipedia DE** page requests per hour)

Proactive Resource Provisioning

Resource provisioning:

(I) Without forecasting (solely reactive):

Resource provisioning actions triggered by **76 SLA violations**

(II) Using WCF approach (proactive):

Reduction to **34 or less SLA violations**

→ No significant change in resource usage observed

(server instances per hour)

8x	correct forecast:	server instance not needed
42 x	correct forecast:	server instance needed at time t
15 x	nearly correct forecast:	time t slightly too early or too late
19 x	incorrect forecast:	need not detected or false positive

day	upper threshold	lower threshold
1		
2		
...		
21		
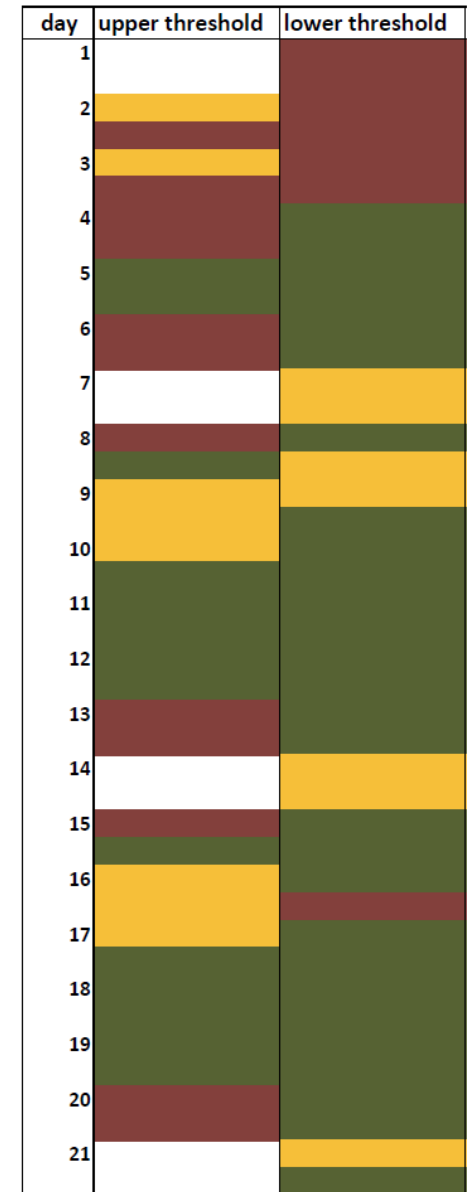
Case Study 2: Automated Derivation of Statistical Regression Models
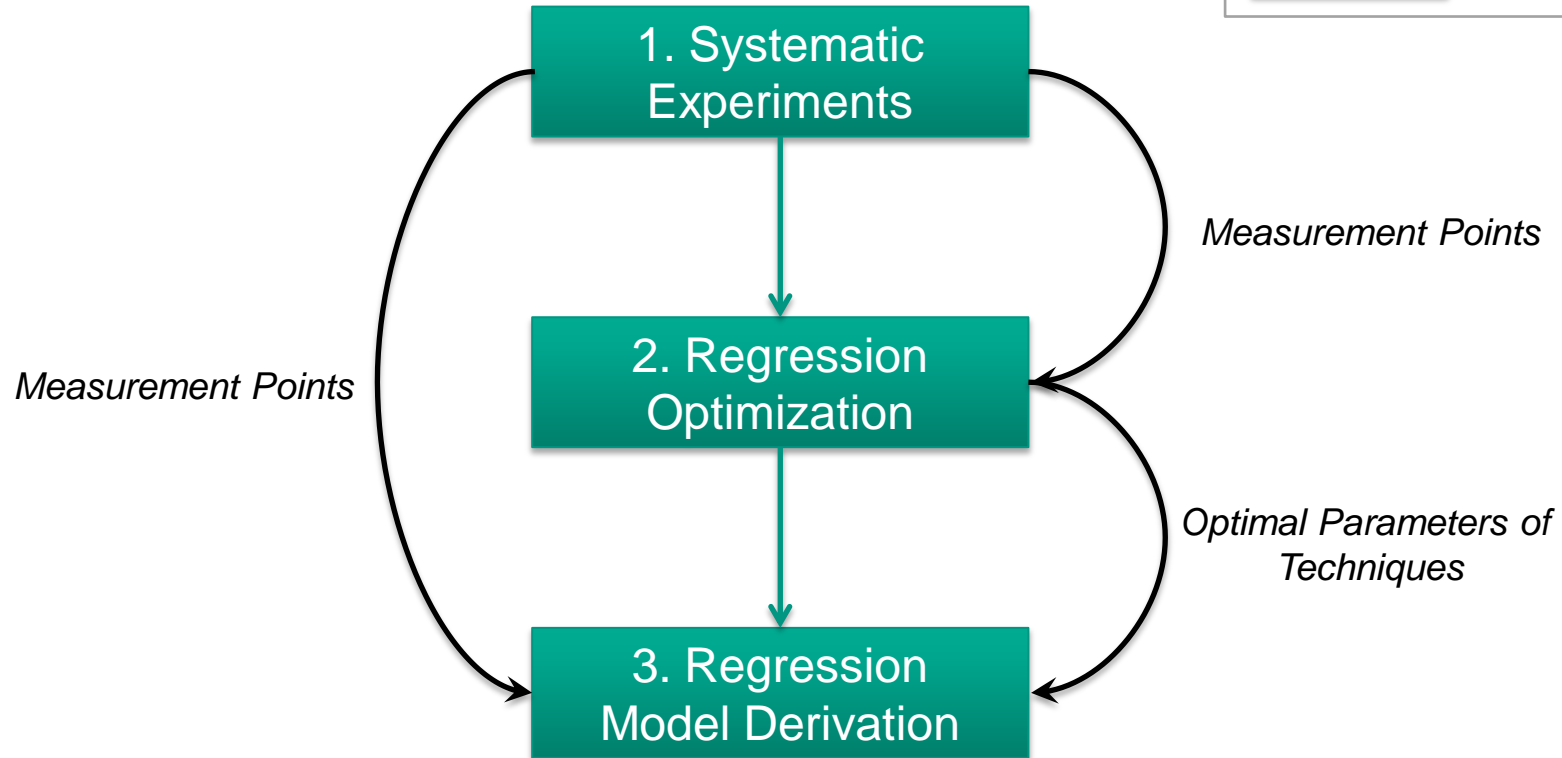
Action — Information Flow

1. Systematic Experiments

2. Regression Optimization

3. Regression Model Derivation

Measurement Points

Measurement Points

Optimal Parameters of Techniques

Q. Noorshams, D. Bruhn, S. Kounev, and R. Reussner. **Predictive Performance Modeling of Virtualized Storage Systems using Optimized Statistical Regression Techniques**. In *Proc. of the 4th ACM/SPEC International Conference on Performance Engineering*, Prague, Czech Republic, ICPE '13, pages 283-294. ACM. 2013. [bib | DOI | http | .pdf]

Example Statistical Regression Models

$$f(x_1, x_2) = c_1 \, x_1 x_2 + c_2 \, x_1 + c_3 \, x_2 + c_4$$

interaction term

Piecewise linearity

LRM

MARS

Tree structure
Step function

CART

Combination

Boosting
Inst.-based

M5

Cubist

Parameters: # of terms, …

Parameters: # of nodes, …

Parameters: # of trees,
# of nodes, …

Parameters: # of nodes, …

LRM - Linear Regression Models
MARS - Multivariate Adaptive Regression Splines
CART - Classification and Regression Trees

Motivation >> Goals & Approach >> **CASE STUDIES** >> Outlook

Online Model-Based Performance and Resource Management in Virtualized Application Environments
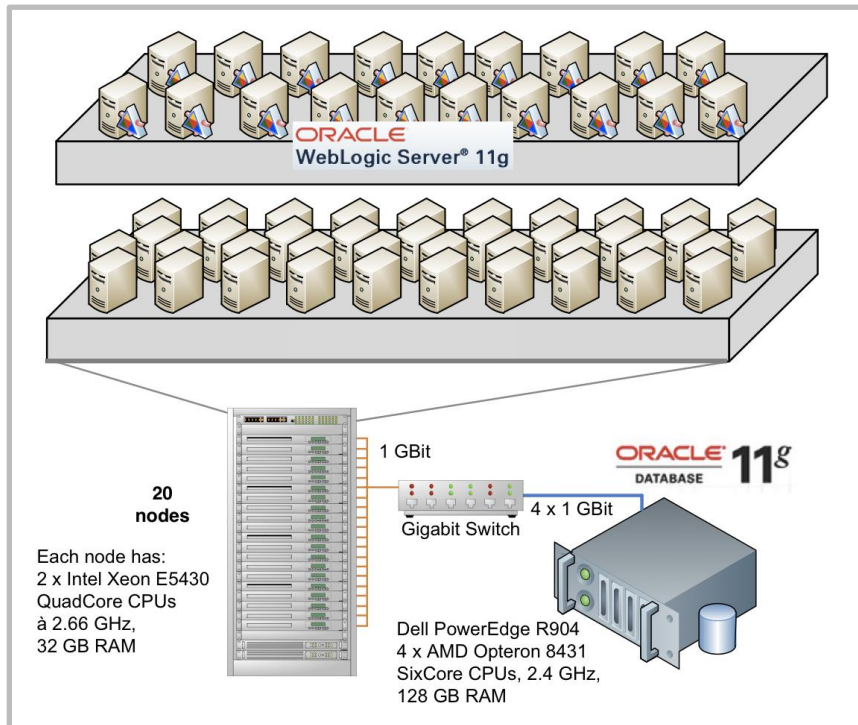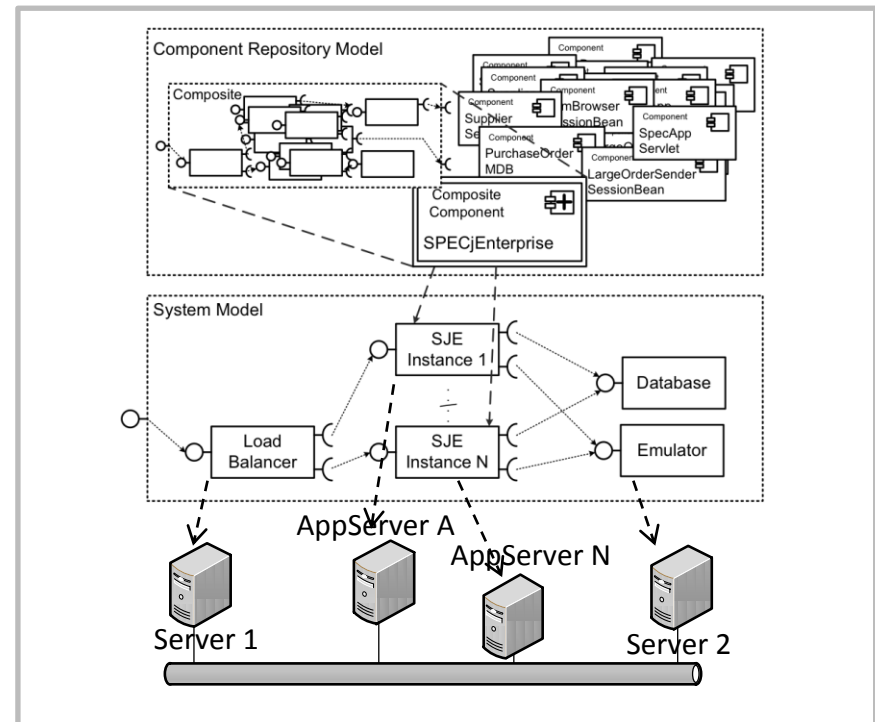
Case Study 3: Automated Extraction of Architecture-level Models

Experimental environment at KIT

- 20 AppServer nodes (20 x 8 core)
- Database server: 24 CPU cores

High-level architecture model overview

- 28 software components
- 63 behavior specifications

F. Brosig, N. Huber, and S. Kounev. **Automated Extraction of Architecture-Level Performance Models of Distributed Component-Based Systems**. In *26th IEEE/ACM International Conference On Automated Software Engineering (ASE 2011)*, Oread, Lawrence, Kansas. November 2011. [bib | .pdf]

Motivation >> Goals & Approach >> **CASE STUDIES** >> Outlook

Project with VMware

- Integration of online performance models into virtualization platforms and virtual appliances

- Model-based performance and resource management
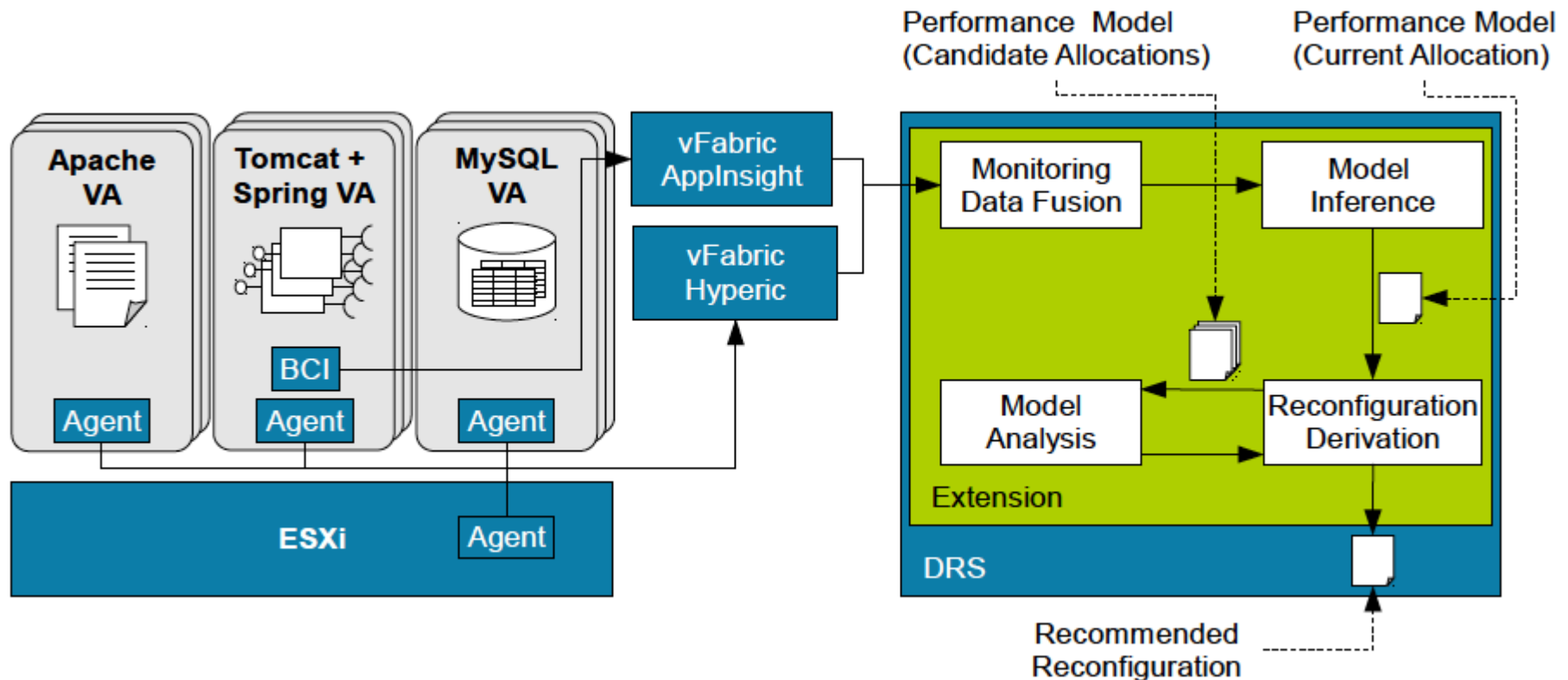
Case Study 4: End-to-End Proof-of-Concept
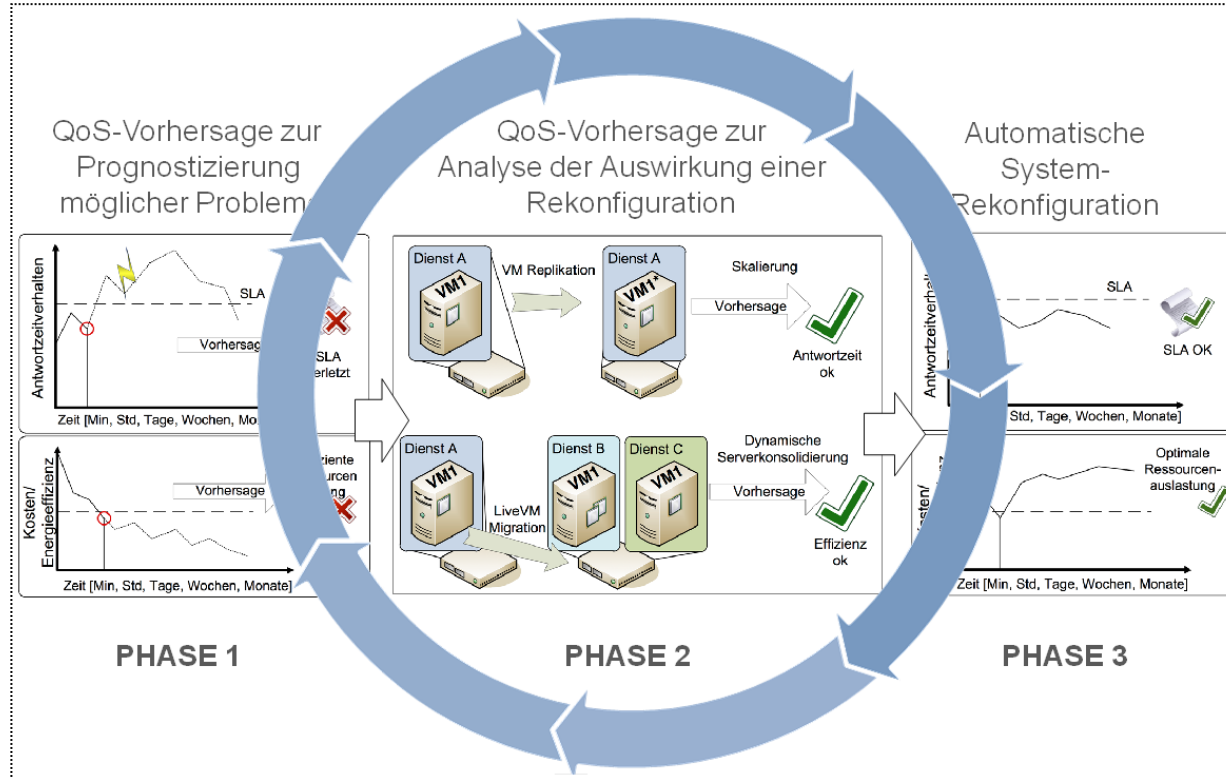
N. Huber, F. Brosig, and S. Kounev. **Model-based Self-Adaptive Resource Allocation in Virtualized Environments**. In *6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2011)*, Waikiki, Honolulu, HI, USA. May 23-24, 2011. [bib | http | .pdf]

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 (SOCA), 2013, Springer. In print. [bib | .pdf]

Modeling Run-time Adaptation Policies

Events / Objectives

trigger / guide

StrategyX   StrategyY

use

TacticA   TacticB   TacticC

execute

Action1 Action2 Action3 Action4 ActionN

reconfigure

System Model /
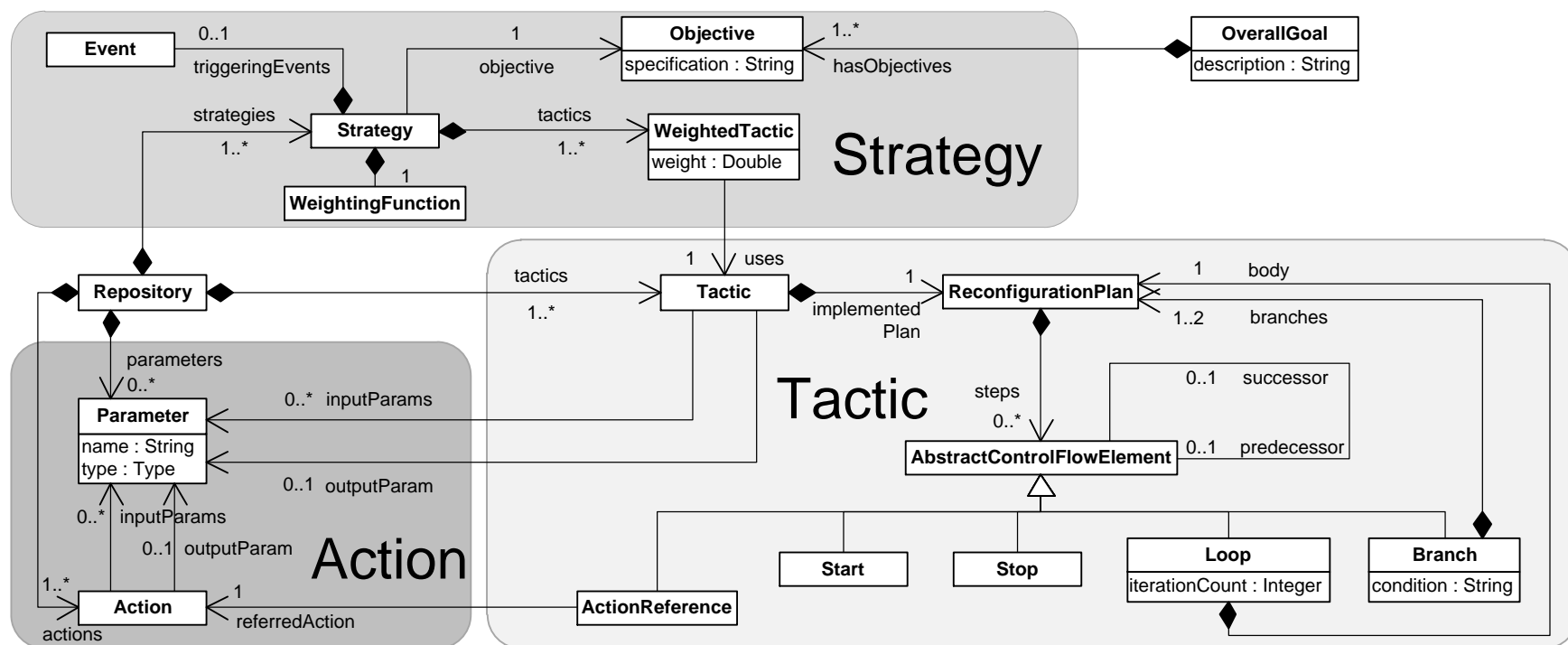Real System
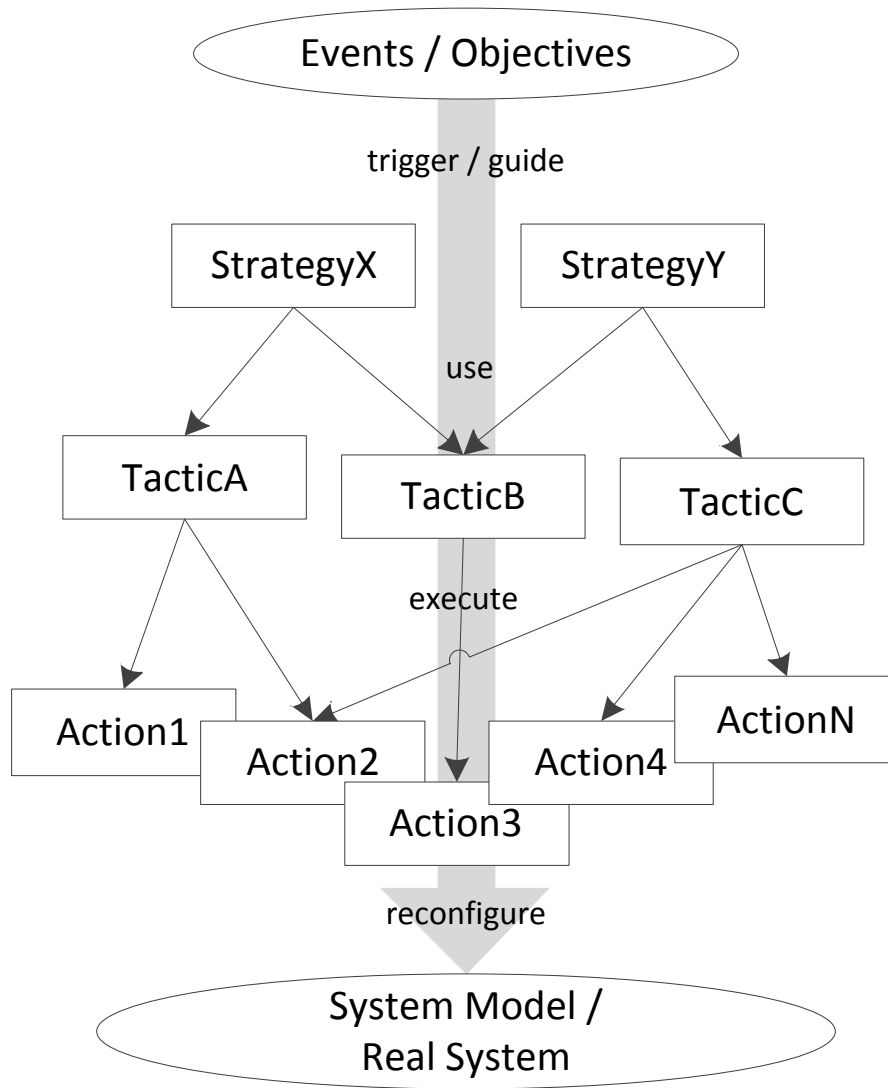
Separate

■ Logical view, high-level process

■ Technical view, low-level operations

[ICEBE 2012, SOCA 2013]

S/T/A Meta-Model

- Actions refer to configuration points
- Tactics execute Actions in Reconfiguration Plans
- Strategies use weighted Tactics

Online Model-Based Performance and Resource Management in Virtualized Application Environments

Further Topics

- **Modeling virtualization** performance **overhead** (CLOSER'11, MASCOTS'13)

- **Performance isolation** in shared execution environments with focus on **multi-tenant SaaS** (QoSA'12, SciCo'13)

- Modeling of **virtualized storage systems** using optimized statistical regression techniques (ICPE'13, MASCOTS' 13)

- **Cloud usage patterns**: A formalism for description of cloud usage scenarios (SPEC Research Group)

- **Intrusion detection and prevention** in virtualized environments (SPEC Research Group)

References

- **Elastic Capacity Management / Online Workload Forecasting**

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

  - N. Herbst, N. Huber, S. Kounev, and E. Amrehn. Self-Adaptive Workload Classification and Forecasting for Proactive Resource Provisioning. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE 2013), Prague, Czech Republic, April 21-24*. 2013. [bib | slides | .pdf]

  - N. Huber, F. Brosig, and S. Kounev. Model-based Self-Adaptive Resource Allocation in Virtualized Environments. In *6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2011)*, Waikiki, Honolulu, HI, USA. May 23-24, 2011. [bib | http | .pdf]

- **Automatic Model Extraction based on Benchmarking or Online System Monitoring**

  - Q. Noorshams, D. Bruhn, S. Kounev, and R. Reussner. Predictive Performance Modeling of Virtualized Storage Systems using Optimized Statistical Regression Techniques. In *Proc. of the 4th ACM/SPEC International Conference on Performance Engineering*, Prague, Czech Republic, ICPE '13, pages 283-294, New York, NY, USA. ACM. 2013. [bib | DOI | http | .pdf]

  - F. Brosig, N. Huber, and S. Kounev. Automated Extraction of Architecture-Level Performance Models of Distributed Component-Based Systems. In *26th IEEE/ACM International Conference On Automated Software Engineering (ASE 2011)*, Oread, Lawrence, Kansas. November 2011. [bib | .pdf]

  - S. Kounev, K. Bender, F. Brosig, N. Huber, and R. Okamoto. Automated Simulation-Based Capacity Planning for Enterprise Data Fabrics. In *4th International ICST Conference on Simulation Tools and Techniques (SIMUTools 2011)*, Barcelona, Spain. March 21-25, 2011. **Best Paper Award**. [bib | slides | .pdf]

  - Q. Noorshams, K. Rostami, S. Kounev, P. Tůma, and R. Reussner. I/O Performance Modeling of Virtualized Storage Systems. In *Proceedings of the IEEE 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, San Francisco, USA. 2013. [bib | .pdf]

- **Performance Modeling and Prediction in Virtualized Environments**

  - F. Brosig, F. Gorsler, N. Huber, and S. Kounev. Evaluating Approaches for Performance Prediction in Virtualized Environments. In *Proceedings of the IEEE 21st International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, San Francisco, USA. 2013. [bib | .pdf]

  - N. Huber, M. von Quast, M. Hauck, and S. Kounev. Evaluating and Modeling Virtualization Performance Overhead for Cloud Environments. In *Proceedings of the International Conference on Cloud Computing and Services Science (CLOSER 2011)*, Noordwijkerhout, The Netherlands, pages 563 - 573. SciTePress. May 7-9, 2011. **Best Paper Award**. [bib | http | .pdf]

References

- **Descartes Meta-Model (DMM) - Online Models for Architecting Self-Aware Systems**

  - http://www.descartes-research.net/research_and_profile/descartes_meta_model/

  - F. Brosig, N. Huber, and S. Kounev. Architecture-Level Software Performance Abstractions for Online Performance Prediction. *Elsevier Science of Computer Programming Journal (SciCo)*, 2013. [bib | DOI | .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 (SOCA), 2013, Springer. In print. [bib | .pdf]

  - N. Huber, A. van Hoorn, A. Koziolek, F. Brosig, and S. Kounev. S/T/A: Meta-Modeling Run-Time Adaptation in Component-Based System Architectures. In *9th IEEE Intl. Conf. on e-Business Engineering (ICEBE 2012), Hangzhou, China*, September 9-11, 2012. [bib | http | .pdf]

  - F. Brosig, N. Huber, and S. Kounev. Modeling Parameter and Context Dependencies in Online Architecture-Level Performance Models. In Proc. of the 15th ACM SIGSOFT Intl. Symposium on Component Based Software Engineering (CBSE 2012), June 26-28, 2012. [bib | .pdf]

  - N. Huber, F. Brosig, and S. Kounev. Modeling Dynamic Virtualized Resource Landscapes. In Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012), Bertinoro, Italy, June 25-28, 2012. [bib| .pdf]

  - S. Kounev, F. Brosig, and N. Huber. Descartes Meta-Model (DMM). Technical report, Karlsruhe Institute of Technology (KIT), 2013. To appear. [bib | http]

- **Vision of Self-Aware Computing Systems**

  - "Model-driven Algorithms and Architectures for Self-Aware Computing Systems" Dagstuhl Seminar scheduled to take place in October 2014 organized by Samuel Kounev, Jeff Kephart, Marta Kwiatkowska and Xiaoyun Zhu.

  - S. Kounev. Engineering of Self-Aware IT Systems and Services: State-of-the-Art and Research Challenges. In *Proc. of the 8th European Performance Engineering Workshop (EPEW'11), Borrowdale, The English Lake District, October 12-13*. 2011. (Keynote Talk). [bib | .pdf]

  - S. Kounev. Self-Aware Software and Systems Engineering: A Vision and Research Roadmap. In *GI Softwaretechnik-Trends, 31(4), November 2011, ISSN 0720-8928*, Karlsruhe, Germany, 2011. [bib | .html | .pdf]

  - S. Kounev, F. Brosig, and N. Huber. Towards self-aware performance and resource management in modern service-oriented systems. In *Proc. of the 7th IEEE Intl. Conference on Services Computing (SCC 2010), July 5-10, Miami, Florida, USA*. IEEE, 2010. [bib | .pdf]

References

- Cloud Usage Scenarios, Challenges and Opportunities

  - A. Milenkoski, A. Iosup, S. Kounev, K. Sachs, P. Rygielski, J. Ding, W. Cirne, and F. Rosenberg. Cloud Usage Patterns: A Formalism for Description of Cloud Usage Scenarios. Technical Report SPEC-RG-2013-001 v.1.0.1, SPEC Research Group - Cloud Working Group, Standard Performance Evaluation Corporation (SPEC), April 2013. [bib | .pdf]

  - S. Kounev, P. Reinecke, F. Brosig, J. T. Bradley, K. Joshi, V. Babka, A. Stefanek, and S. Gilmore. Providing dependability and resilience in the cloud: Challenges and opportunities. In K. Wolter, A. Avritzer, M. Vieira, and A. van Moorsel, editors, *Resilience Assessment and Evaluation of Computing Systems*, XVIII. Springer-Verlag, Berlin, Heidelberg, 2012. ISBN: 978-3-642-29031-2. [bib | http | .pdf]

- Performance Isolation in Shared Execution Environments (e.g., Multi-Tenant SaaS)

  - R. Krebs, C. Momm, and S. Kounev. Metrics and Techniques for Quantifying Performance Isolation in Cloud Environments. *Elsevier Science of Computer Programming Journal (SciCo)*, 2013. To appear. [bib]

  - R. Krebs, C. Momm, and S. Kounev. Metrics and Techniques for Quantifying Performance Isolation in Cloud Environments. In Barbora Buhnova and Antonio Vallecillo, editors, *Proceedings of the 8th ACM SIGSOFT International Conference on the Quality of Software Architectures (QoSA 2012)*, Bertinoro, Italy, pages 91-100, New York, USA. ACM Press. June 25-28, 2012. [bib | http | .pdf]

- Intrusion Detection and Prevention in Virtualized Environments

  - A. Milenkoski, S. Kounev, A. Avritzer, N. Antunes, and M. Vieira. On Benchmarking Intrusion Detection Systems in Virtualized Environments. Technical Report SPEC-RG-2013-002 v.1.0, SPEC Research Group - IDS Benchmarking Working Group, Standard Performance Evaluation Corporation (SPEC), June 2013. [bib | .pdf]

Thank You!

http://www.descartes-research.net

Dagstuhl Seminar "Model-driven Algorithms and Architectures for Self-Aware Computing Systems" (2014) co-organized by S. Kounev (KIT), J. Kephart (IBM), M. Kwiatkowska (Oxford) and X. Zhu (VMware).

http://www.linkedin.com/groups/SelfAware-Computing-5103054

 Online Model-Based Performance and Resource Management in Virtualized Application Environments