Performance & Availability & Power Quantification

for Infrastructure-as-a-Service Cloud

Kishor Trivedi

kst@ee.duke.edu

www.ee.duke.edu/~kst

Dept. of ECE, Duke University, Durham, NC 27705

NSF PI Meeting, Washington DC

March 17-18, 2011

Summary of Research

- **Performance Quantification of IaaS Cloud**
  - Scalable performance evaluation using interacting stochastic models [PRDC 2010]
- **Availability Quantification of IaaS Cloud**
  - Scalable availability evaluation using interacting stochastic models [DSN 2011]
- **Power Consumption Quantification in IaaS Cloud**
  - Trade-offs in performance and power consumption [DSN 2011 workshop]
- **Future Research**

Key Challenges

- Three critical metrics for a cloud:
  - Service (un)availability
  - Performance (response time) unpredictability
  - Power consumption

- Large number of parameters can affect performance, availability and power
  - Workload parameters
  - Failure/recovery characteristics
  - Types of physical infrastructure
  - Characteristics of virtualization infrastructures
  - Large scale; thousands of servers

Performance & availability & Power quantification is difficult!

Our Goals in the IBM Clouds Project

- Develop a comprehensive analytic modeling approach
- High fidelity
- Scalable and tractable
- Apply these models to cloud capacity planning

Joint work with:
Rahul Ghosh and Dong Seong Kim (Duke), Francesco Longo (Univ. of Messina),
Vijay Naik, Murthy Devarakonda and Daniel Dias
(IBM T. J. Watson Research Center)

Performance and Availability Analysis

- Difficulty with measurement-based approach:
  - expensive experimentation for each workload and system configuration

- Monolithic analytic model will suffer largeness and hence not scalable

- Our approach:
  - sub-model solutions composed via an interacting Markov chain approach
  - scalable and tractable
  - low cost of model solution while covering large parameter space

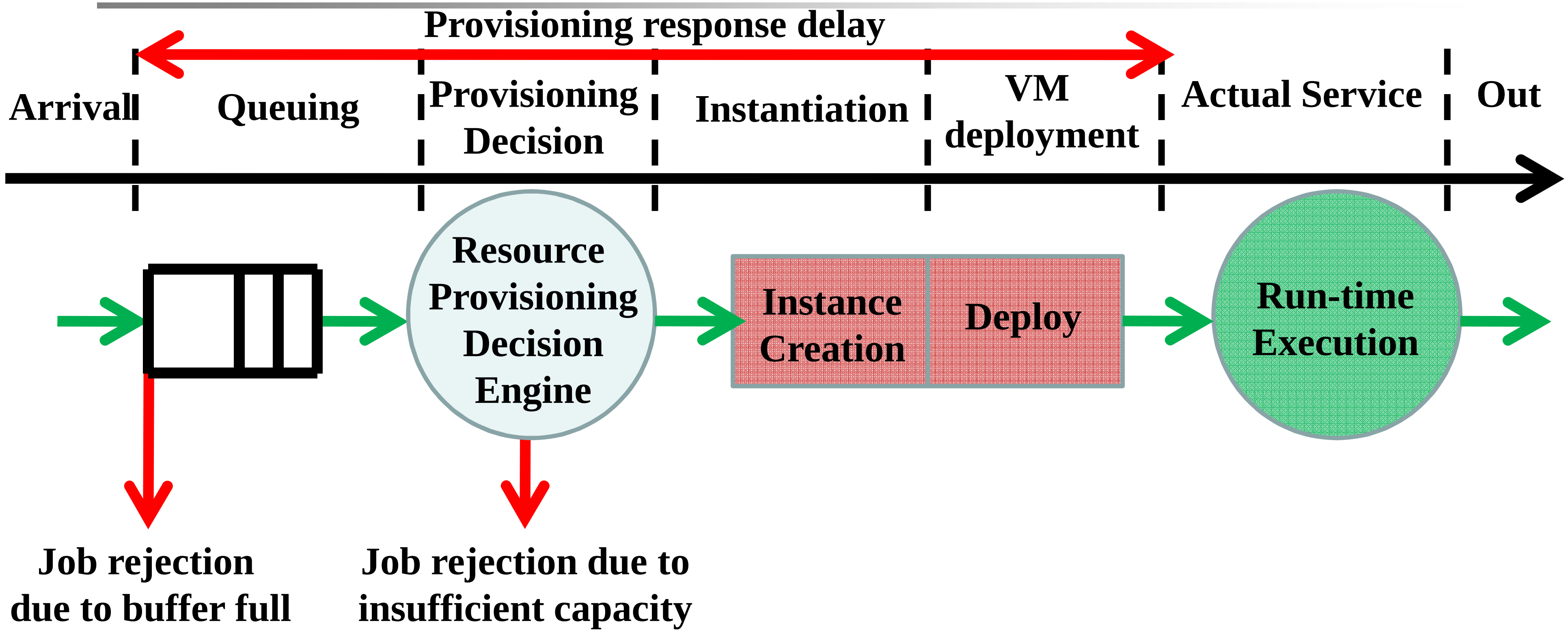
System model

- **Current Assumptions [will be relaxed soon]**
  - Homogenous requests
  - All physical machines (PMs) are identical.

To minimize power consumption, PMs divided into three pools:

- **Hot pool– fast provisioning but high power usage**
- **Warm pool—slower provisioning but lower power usage**
- **Cold pool – slowest provisioning but lowest power usage**

Life-cycle of a job inside a IaaS cloud



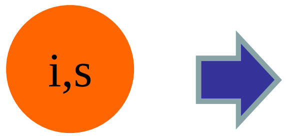
- **Provisioning and servicing steps:**

(i) resource provisioning decision,

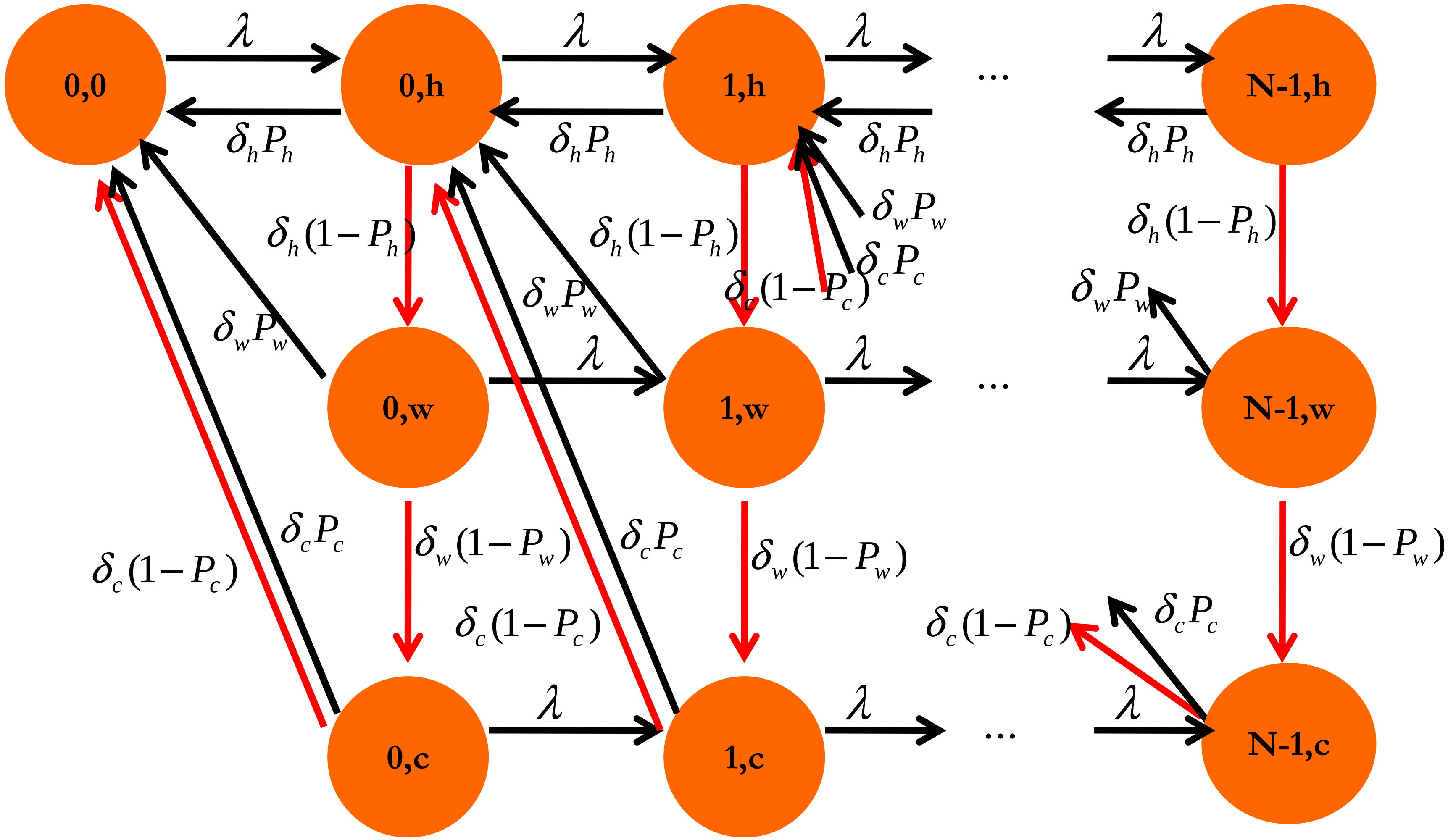
(ii) VM provisioning

We translate these steps into analytic sub-models

Resource provisioning decision model: CTMC

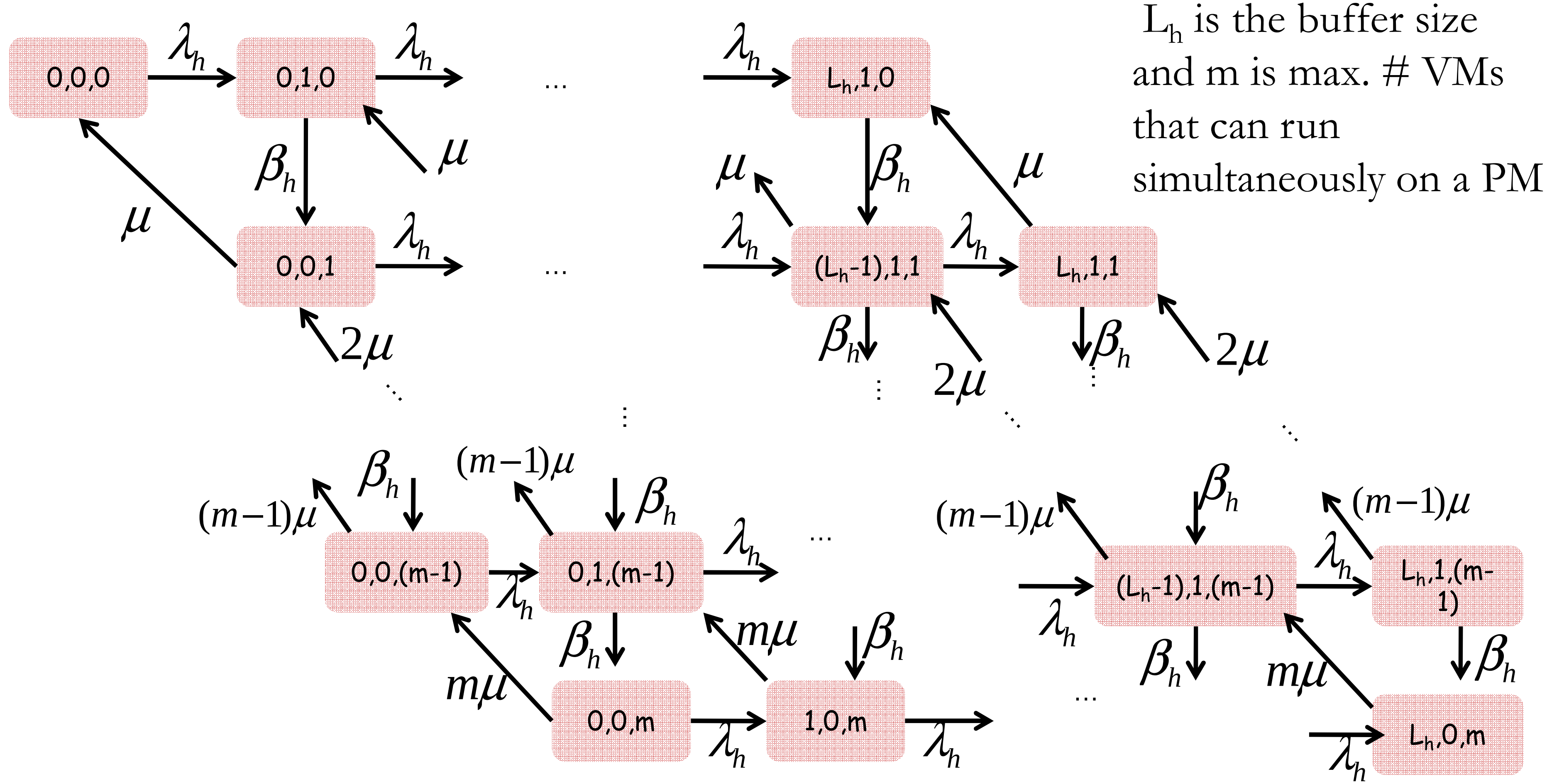


i = number of jobs in queue,
s = pool (hot, warm or cold)



VM provisioning model for each hot PM

L_h is the buffer size and m is max. # VMs that can run simultaneously on a PM



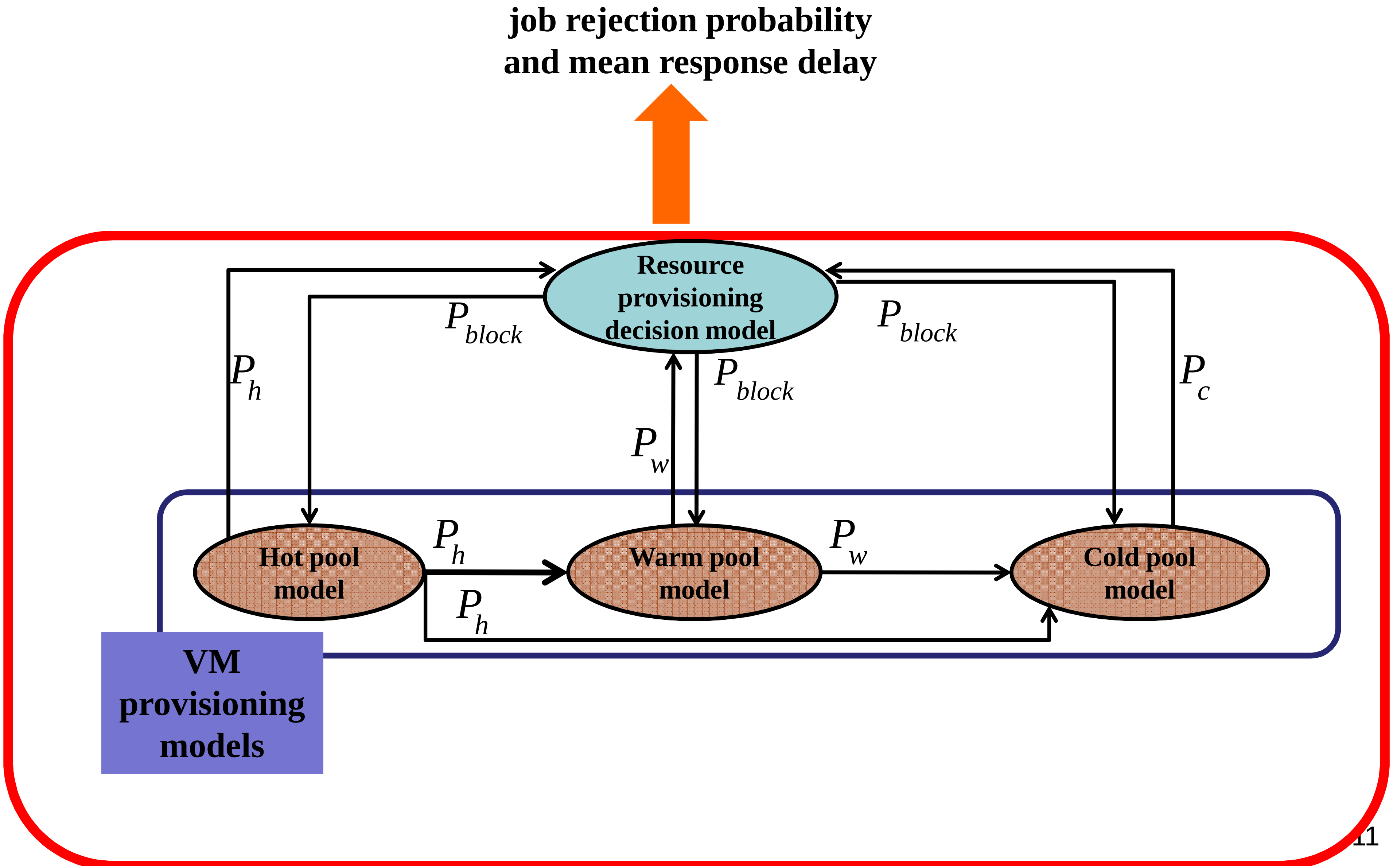
i,j,k → i = number of jobs in the queue, j = number of VMs being provisioned, k = number of VMs running

VM provisioning model: Summary

- Warm/cold PM model is similar to hot PM, except:
  - (i) Effective job arrival rate
  - (ii) For first job, warm/cold PM requires additional start-up time
  - (iii) Mean provisioning delay for a VM for the first job is longer

- Outputs of hot, warm and cold pool models:
  - Probabilities (P_h, P_w, P_c) that at least one PM in hot/warm/cold pool can accept a job

Import graph for performance models



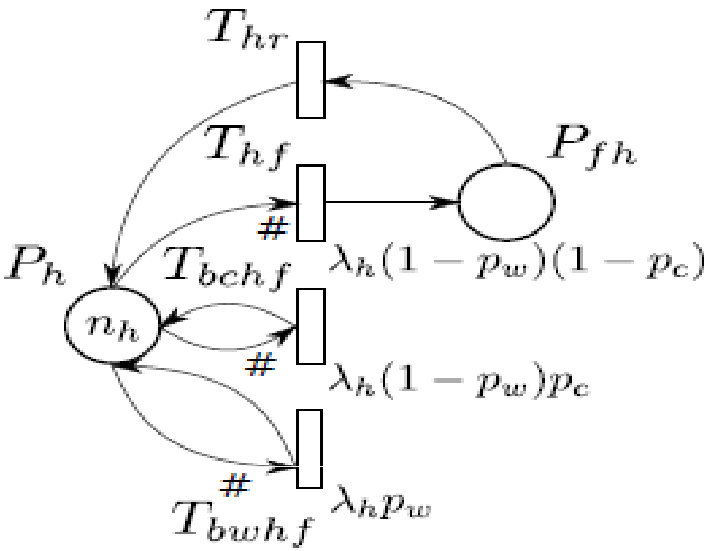
Performance measures comparison with monolithic model

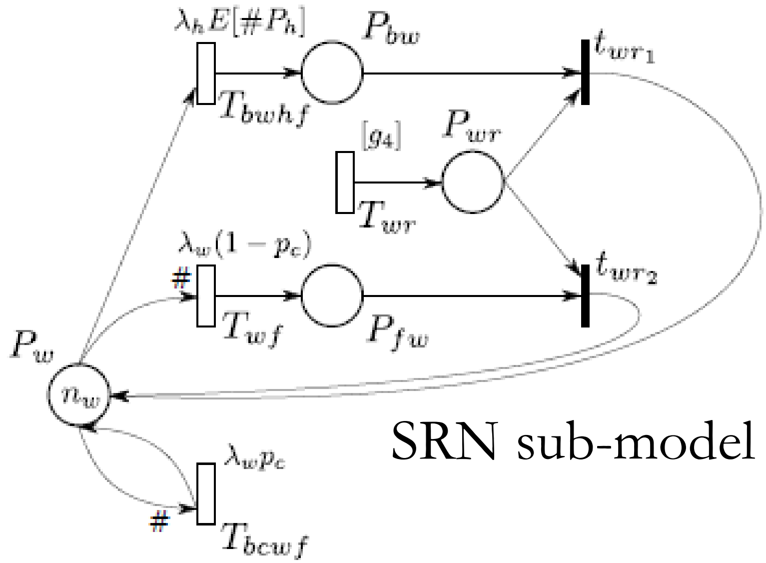
- 1 PM per pool and 1 VM per PM

Jobs/hr	Mean RPDE queue length		Rejection probability	
	ISP	monolithic	ISP	Monolithic
1	9.03319140e-07	9.23205352e-07	9.88988685e-06	1.12206538e-03
5	4.16229278e-05	4.33642749e-05	4.23338652e-02	8.04997306e-02
10	2.37305728e-04	2.42254461e-04	2.34964832e-01	2.65869302e-01
15	6.35391266e-04	6.43770670e-04	3.98599617e-01	4.14929892e-01
20	1.25263680e-03	1.26553370e-03	5.10690667e-01	5.19685843e-01
25	2.09903601e-03	2.11792961e-03	5.89152342e-01	5.94489886e-01
30	3.18263414e-03	3.20909188e-01	6.46478256e-01	6.49848949e-01
35	4.51061770e-03	4.54616990e-03	6.89994801e-01	6.92227235e-01

- The error is between e-03 and e-07 for all the results.
- The number of states in monolithic model is 912 while in ISP model it is 21

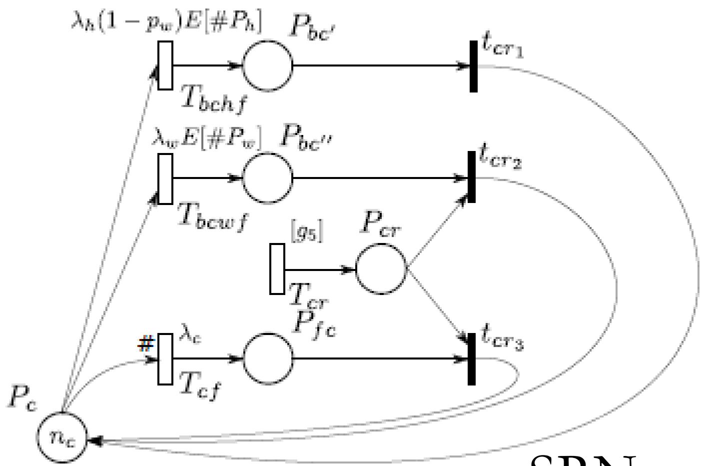
Availability Analysis using Interacting Sub-models





SRN sub-model for warm pool

SRN sub-model for hot pool



SRN sub-model for cold pool

Monolithic vs. ISP Availability Model

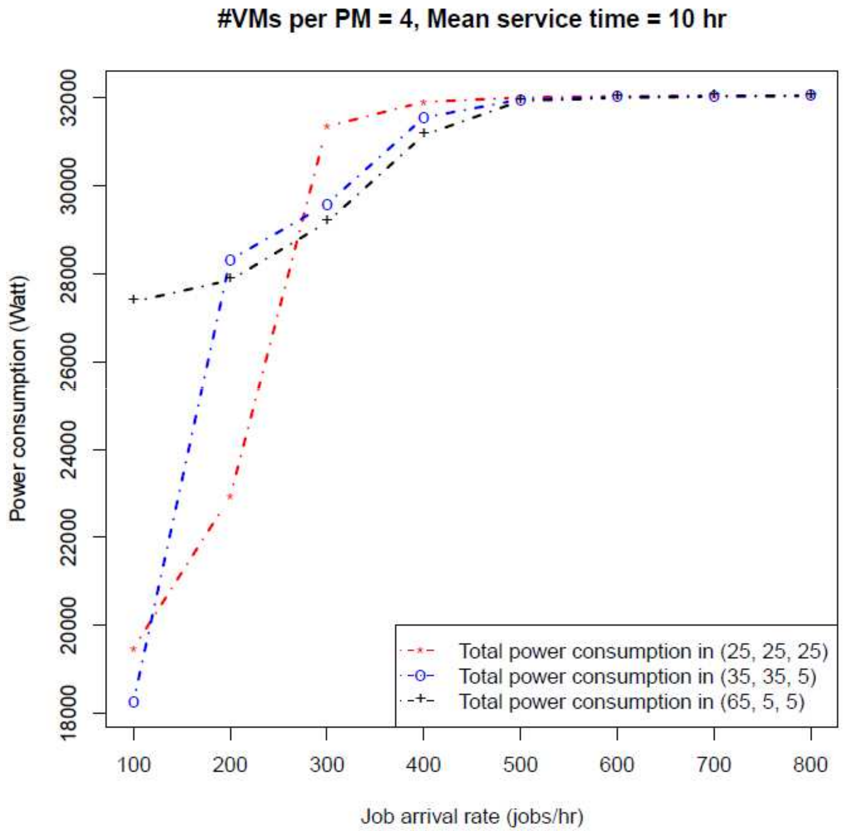
#PMS in each pool	#Monolithic states	#ISP states
5	7056	56
10	207636	286
15	1775616	136
17	3508920	171
19	646800	210
20	Memory overflow	231
50	-	1326
100	-	5151
150	-	11476
200	-	20301

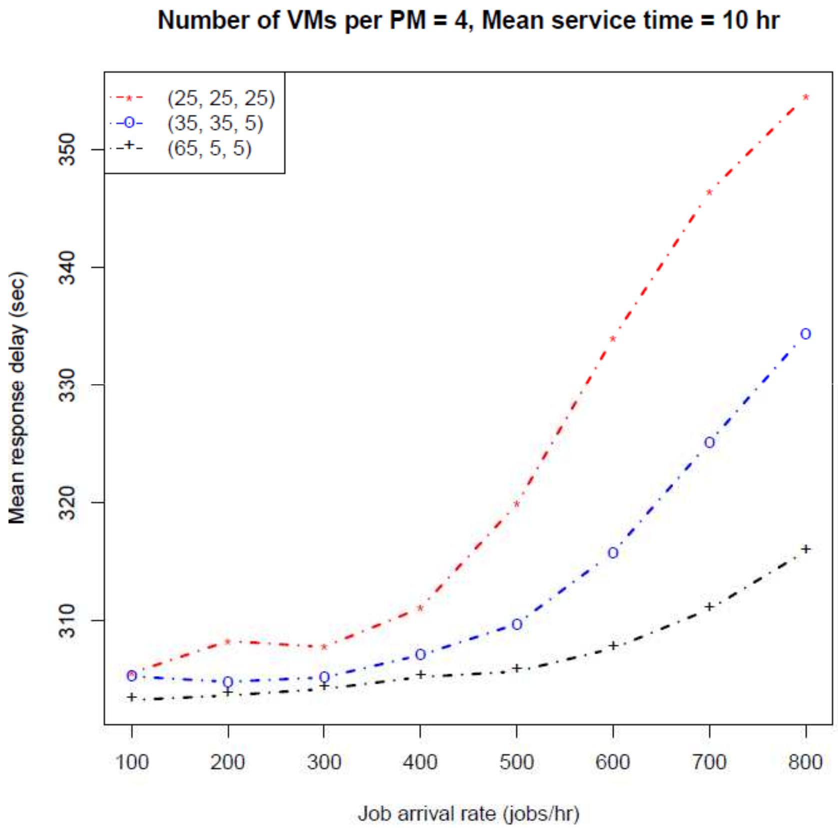
COMPARISON OF SOLUTION TIMES IN SECONDS.

#PMs in each pool in the beginning	Monolithic model	Interacting sub-models
5	0.627	0.406
10	18.670	0.517
15	373.822	0.278
17	1004.494	0.279
19	2459.553	0.280
20	Memory overflow	0.281
50	-	0.296
100	-	0.377
150	-	0.564
200	-	0.948

Power Consumption Analysis

- Markov reward approach used to compute power from Performance model





optimization problem:
What is the optimal #PMs per pool that minimizes total power consumption but does not violate the SLA (upper bound on mean response delay)?

Extensions to Current Models

- Different workload arrival processes [e.g., bursty]
- Different types of service time distributions
- Heterogeneous requests
- Requests with different priorities
- Availability model with all h/w and s/w [e.g., Mandelbugs] failure modes
- Model parameterization from real data and model validation
- Use these models for cloud capacity planning

Conclusions

- Analytic models are powerful for the construction and numerical solution of various reliability, availability, performance, survivability and resiliency [behavior under changes in workload, faultload, configuration] models.
- Not only exponential but also non-exponential distribution can be admitted to construct such models
- For very complex systems such as clouds, hierarchical, fixed-point iterative and approximate solutions needed
- Simulative and hybrid models/solutions should be used only when absolutely necessary
- Parameterization and Validation crucial
- Models can then be used in
  - capacity planning
  - feedback control setting for adapting to changes