Automatic Energy Efficiency Management of Data Center Servers Operated in Hot and Cold Standby and with Dynamic Voltage and Frequency Scaling (DVFS)

Paul J. Kühn
University of Stuttgart, Germany
Institute of Communication Networks and Computer Engineering (IKR)
E-Mail: paul.j.kuehn@ikr.uni-stuttgart.de
Phone: +49-711-685-68027

Dagstuhl Seminar 13131 "Future Internet"
Dagstuhl, March 24 - 27, 2013
and
KuVS/ITG 5.2.4 Working Group Meeting on "Competitive Service Delivery Infrastructures", Vodafone Training Center, Königswinter, April 17, 2013

OUTLINE

1. INFORMATION CENTRIC NETWORKING

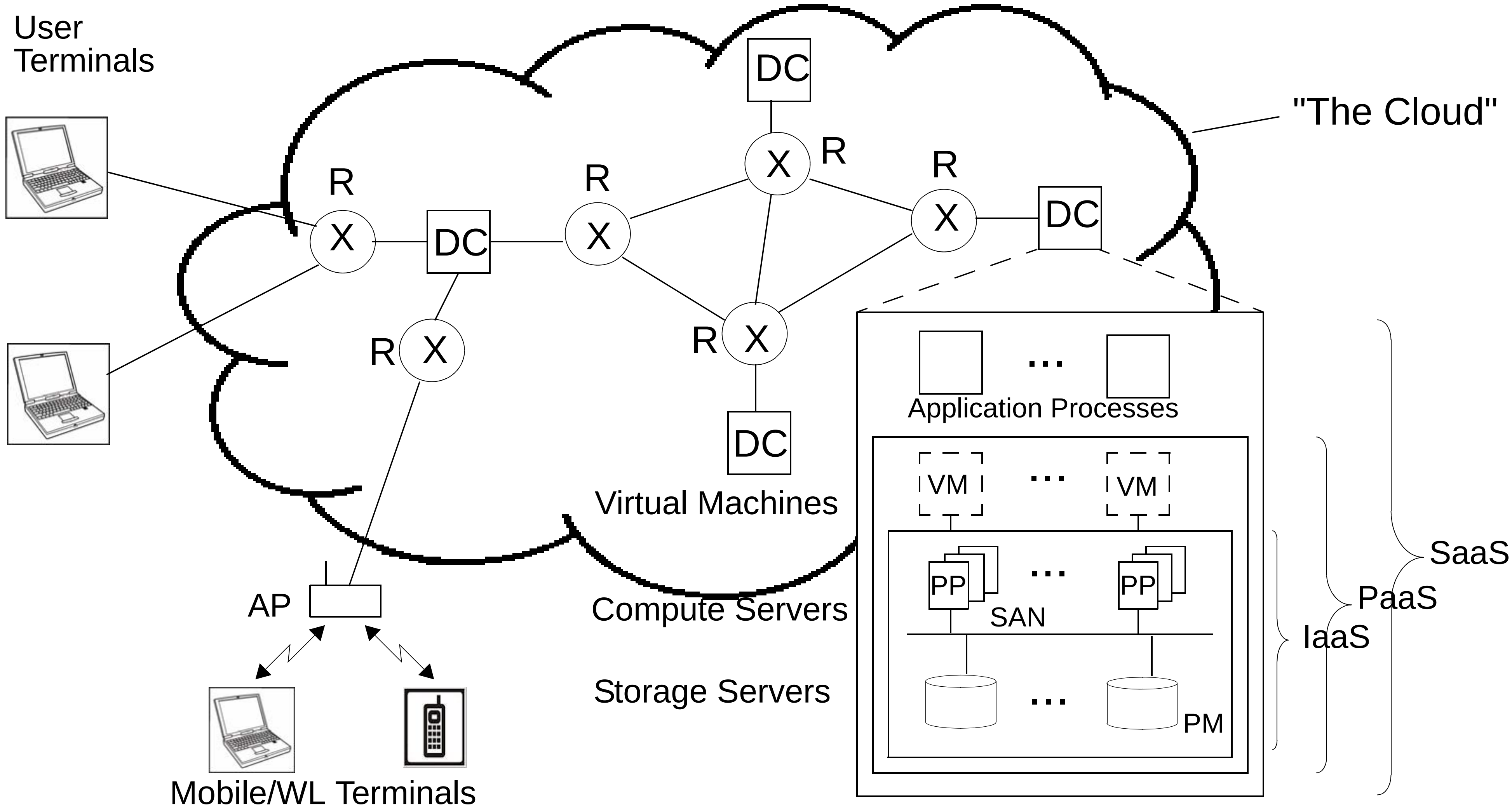
- Major Application Shifts in the Internet

Remote Computation
File Transfer
E-Mail
Content Access (Web)
Peer-to-Peer File Sharing
Voice over IP / Multimedia
Web-based e-Business
Social Networking
Content Access & Distribution

- Paradigm Shifts

Transport Network	----->	Information-Centric Network
Fixed Infrastructure	----->	Wireless and Mobile Infrastructure
End-to-End Control	----->	Network Control
Non-Realtime	----->	Realtime
Best Effort Service	----->	Service-Oriented Network (QoS, QoE, SLA)

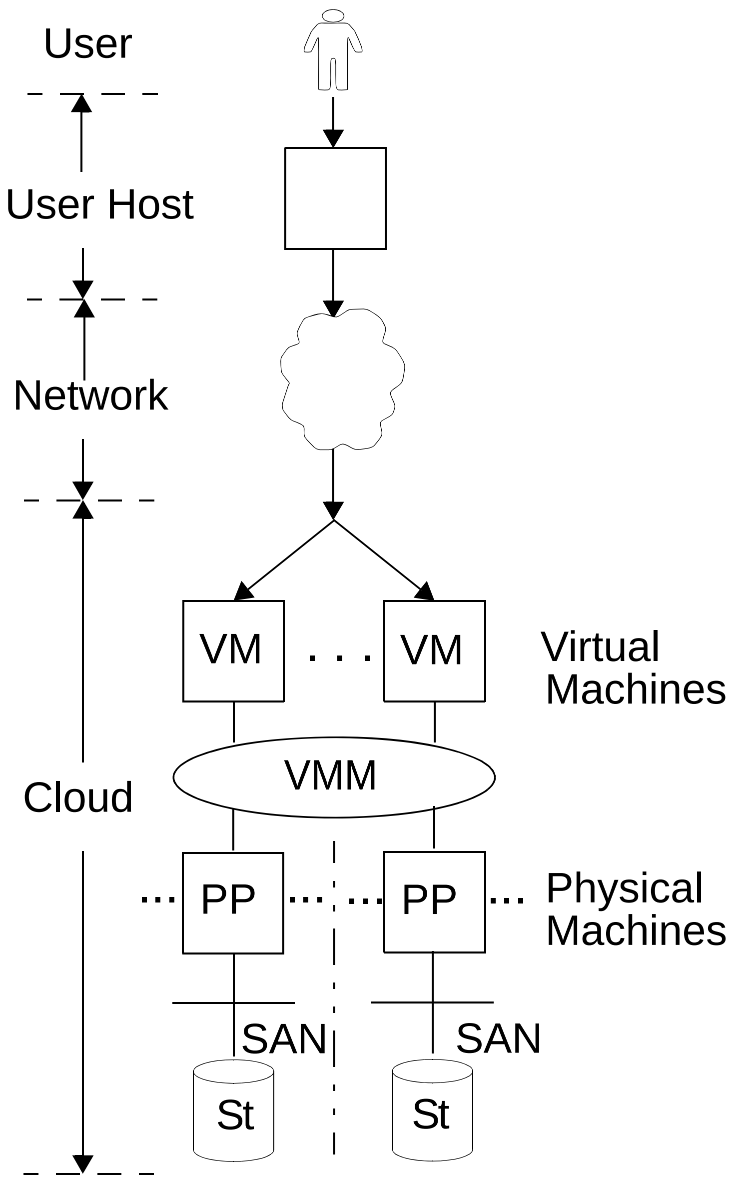
- Current Internet -----> Next Generation / Future Internet


User
Terminals
DC
"The Cloud"
R
X
R
R
R
X
X
DC
X
DC
R
X
R
X
DC
Virtual Machines
Compute Servers
Storage Servers
AP
Mobile/WL Terminals
...
Application Processes
VM
...
VM
PP
...
PP
SAN
...
PM
SaaS
PaaS
IaaS

CLOUD TYPES:	- Public, Private, Hybrid
CLOUD APPLICATIONS:	- Data Retrieval (Web)
	- Content Delivery
	- Business Processes
	- Scientific Grid
	- Social Networking
CLOUD FUNCTIONS:	- Resource Virtualization and Process Migration
	- Resource Sharing
INCENTIVES:	- Economics (Outsourcing/Insourcing of IT Services)
	- Reliability
	- Energy Reduction

2. CONTENT DISTRIBUTION AND CLOUD COMPUTING - RESEARCH ASPECTS

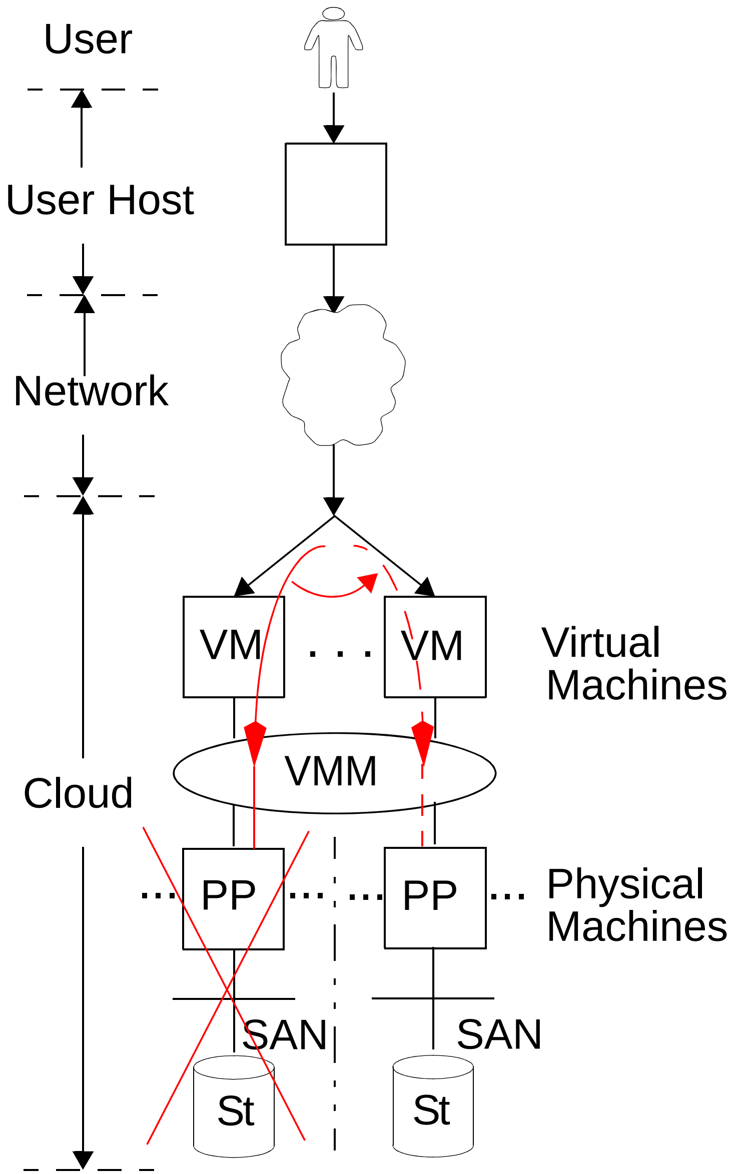
CLOUD ARCHITECTURES:	- Process Migration
	- Operating Systems, Hypervisor
	- Security and Privacy Protection
RESOURCE MANAGEMENT:	- Storage Strategies
	- Scheduling, Routing
	- Admission/Flow/Congestion Control
TRAFFIC ENGINEERING:	- Cloud Traffic Volumes/Characteristics
	- Traffic Matrix, Load Balancing
	- Quality of Service/Experience (QoS/QoE)
ECONOMIC ASPECTS:	- Tradeoff between Storage, Processing, and Communication
	- Service Level Agreements
	- Optimization



Cloud: Pool of Physical Resources Interconnected by Network

VM: Virtual Machine Virtualized View on the Resource Pool

VMM: VM Monitor ("Hypervisor") Mapping of VM to PM



Cloud: Pool of Physical Resources
Interconnected by Network

VM: Virtual Machine
Virtualized View on the Resource Pool

VMM: VM Monitor ("Hypervisor")
Mapping of VM to PM

VM Migration:

- Change of Assignment VM --- PM
- Different Migration Strategies

"Suspend-and-Copy"

"Pre-Copy"

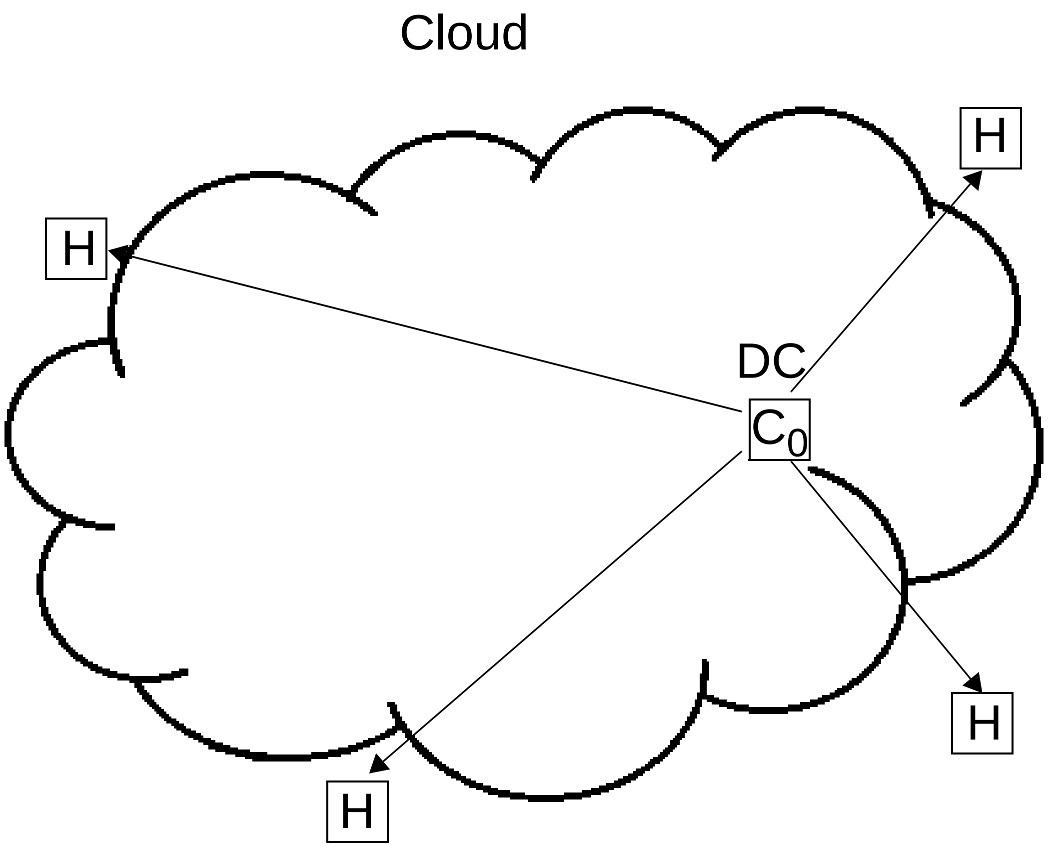
"Post-Copy"

3. MANAGING CDNs - DYNAMIC PROVISIONING OF PHYSICAL RESOURCES

• Incentives	Hot Spot Mitigation	---->	Overload Avoidance
	Load Balancing	---->	Economic Capacity Utilization, Energy Saving
	Server Consolidation	---->	Avoiding "Sprawling" of Resources
	Performance/SLA	---->	Meeting RT Requirements
	Economics	---->	Trade-off between Storage Cost -- Communication Cost in Case of Content Storage Replication

- Content Location: Centralized or Decentralized
- Address Resolution by Publish/Subscribe Mechanism NNC (Network Named Content)
  Translation NNC ----> IP Address
  (Problem of the Legacy Internet without Identifier/Locator Split!)

3. MANAGING CDNs - CENTRALIZED STORAGE

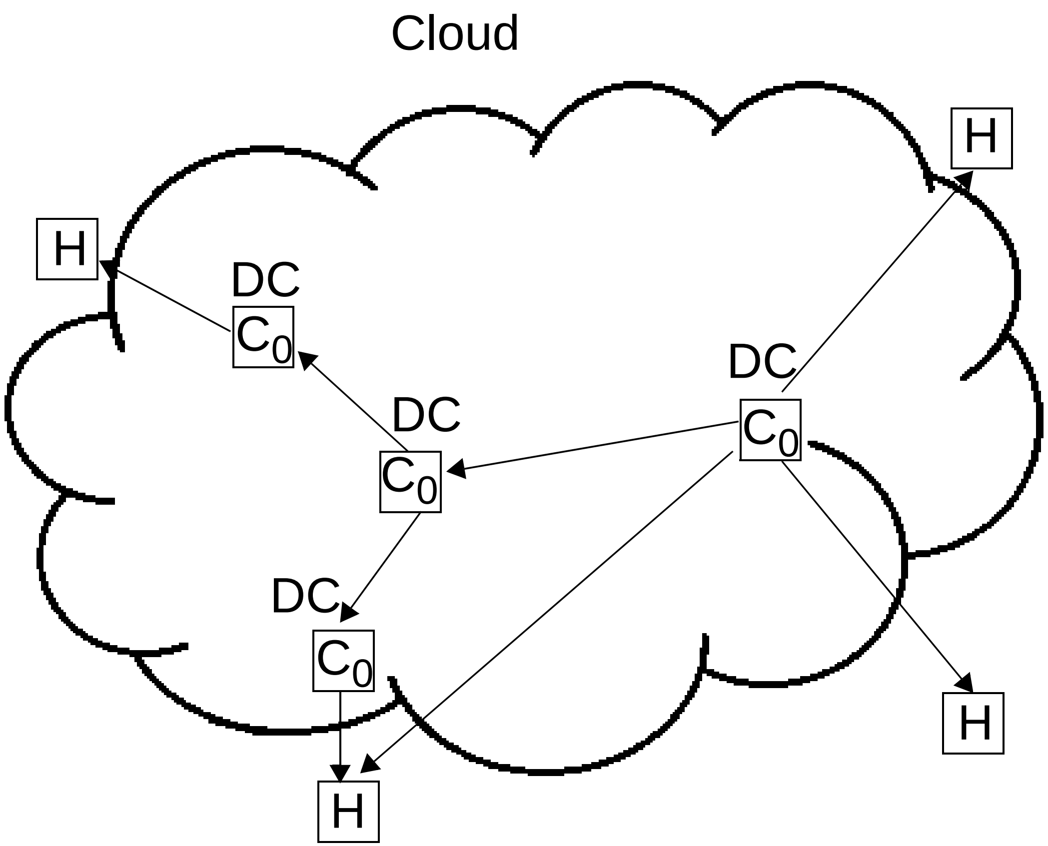


- Multicast Tree
- Minimum Storage Cost
- Maximum Communication Cost
- Maximum Latency
- High Risk, Reliability

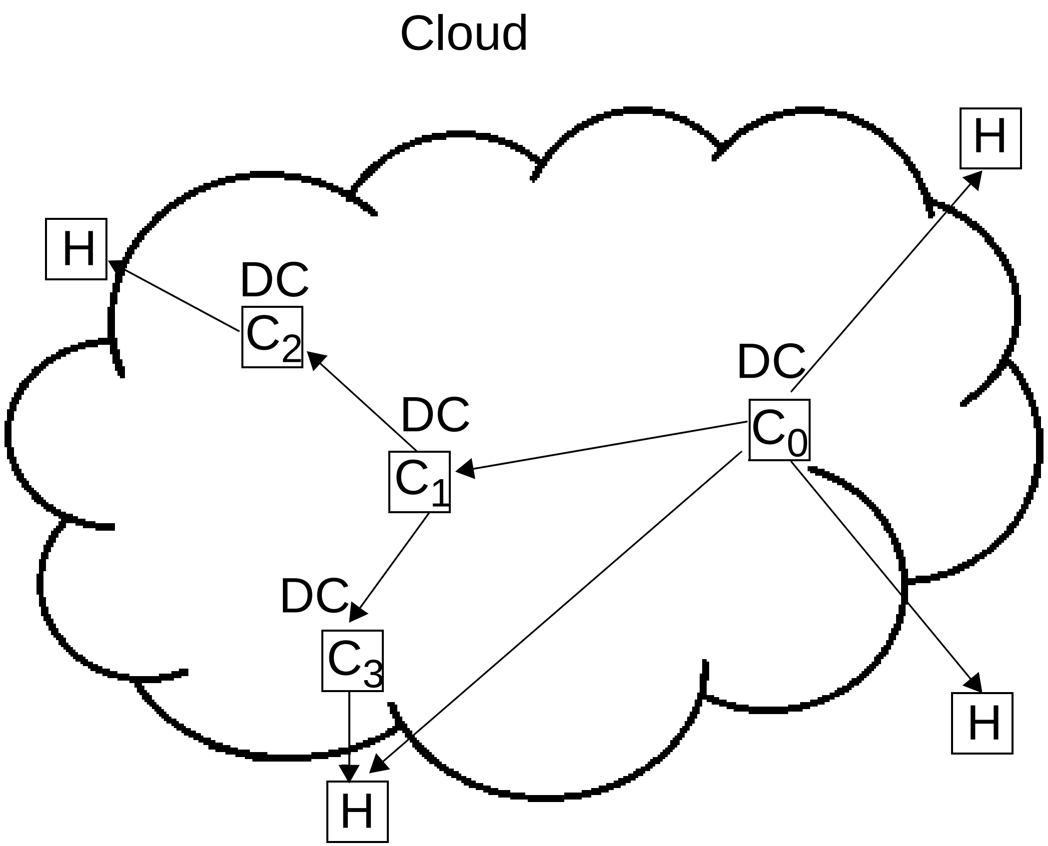
H User Host

DC Data Center

C_0 Content



- Replication of Full Content C_0 by Content Migration
- Higher Storage Cost
  Less Communication Cost
- Short Latency
- Overhead Cost by Replication



- C_0 Full Content
- C_i Partial Content

  $C_i \subseteq C_0$

  $C_2, C_3 \subseteq C_1$
- Dynamic Replication Dependent on Actual Demand
- Replicated Content Storage Management by Caching + LRU Replacement Strategy (Least Recently Used)

Open Questions: Dependence on "Working Set" of Content?
Caching of Content Fragments (Chunks, Packets, whole Objects)?
Amount of Prefetching to Avoid Starvation?
Performance, Energy Demand/Saving?

4. MODELING ALGORITHMS

Modeling Assumptions:

- Cloud with Distributed Data Centers
- NNC Address Resolution by Publish/Subscribe Service
- Multi-Server Model for DC Content Delivery
- Sleep Mode + Activation Delays for Multi-Core Nodes
- Self-Adapting Activation/Deactivation of Core Nodes within each DC (state-dependent; can be extended to Measurement- or Forecast-Based Operation)

4. MODELING ALGORITHMS

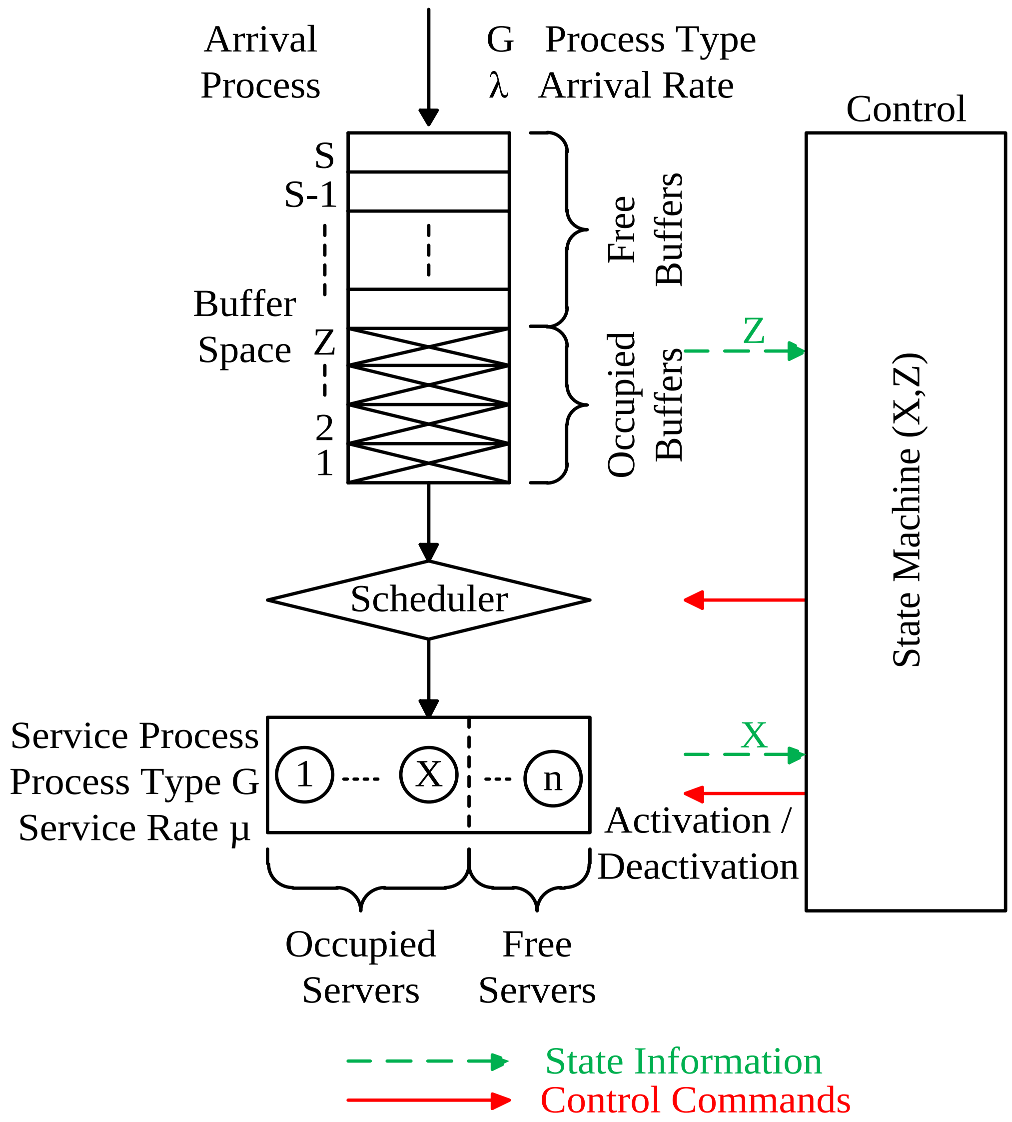
BASIC IDEAS:

- Self-Adapting Operation of Data Center Resources
  - Local Monitoring of Load Development
  - Local Control of Resource Activation/Deactivation by FSM

BASIC MODEL:

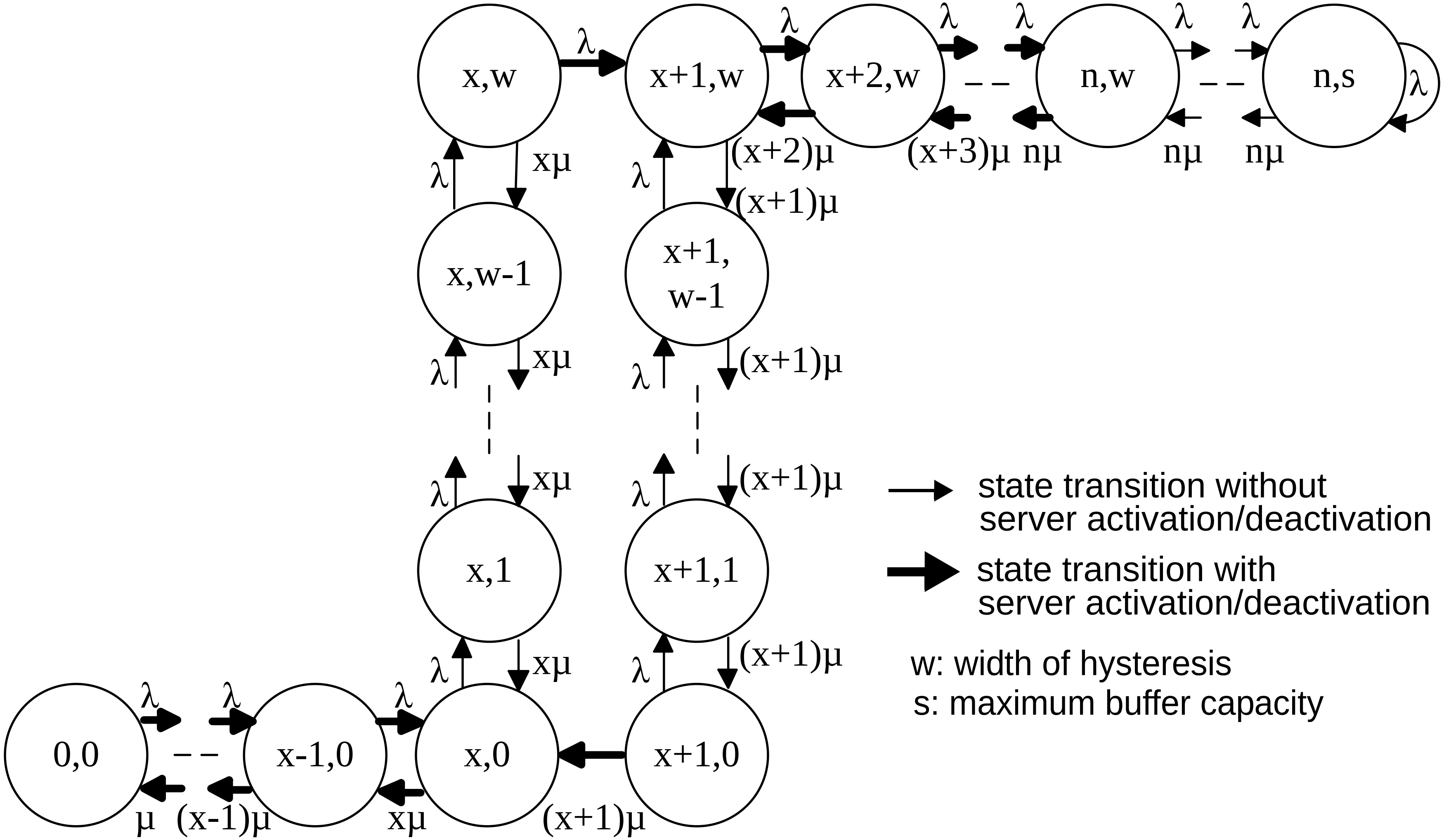
- Uniform Services, N Data Centers
- Focus on Processing Resources only
- (n_i, ρ_i) Resource/Utilization Vector of DC_i, $i \in [1, N]$

INDIVIDUAL DC MODEL



NON-ADAPTING MODEL BY FSM

(1) SINGLE HYSTERESIS MODEL



state transition without server activation/deactivation

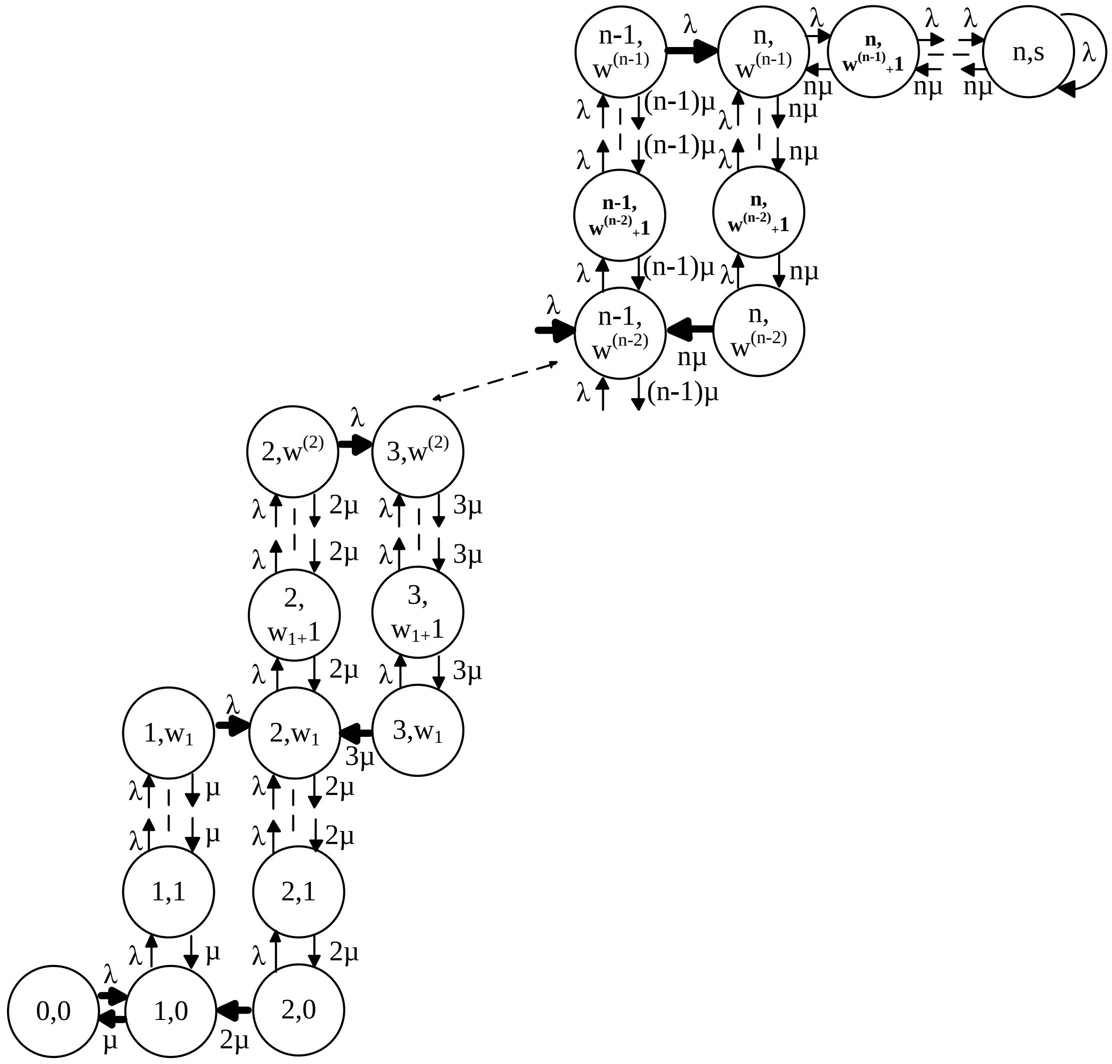
state transition with server activation/deactivation

w: width of hysteresis

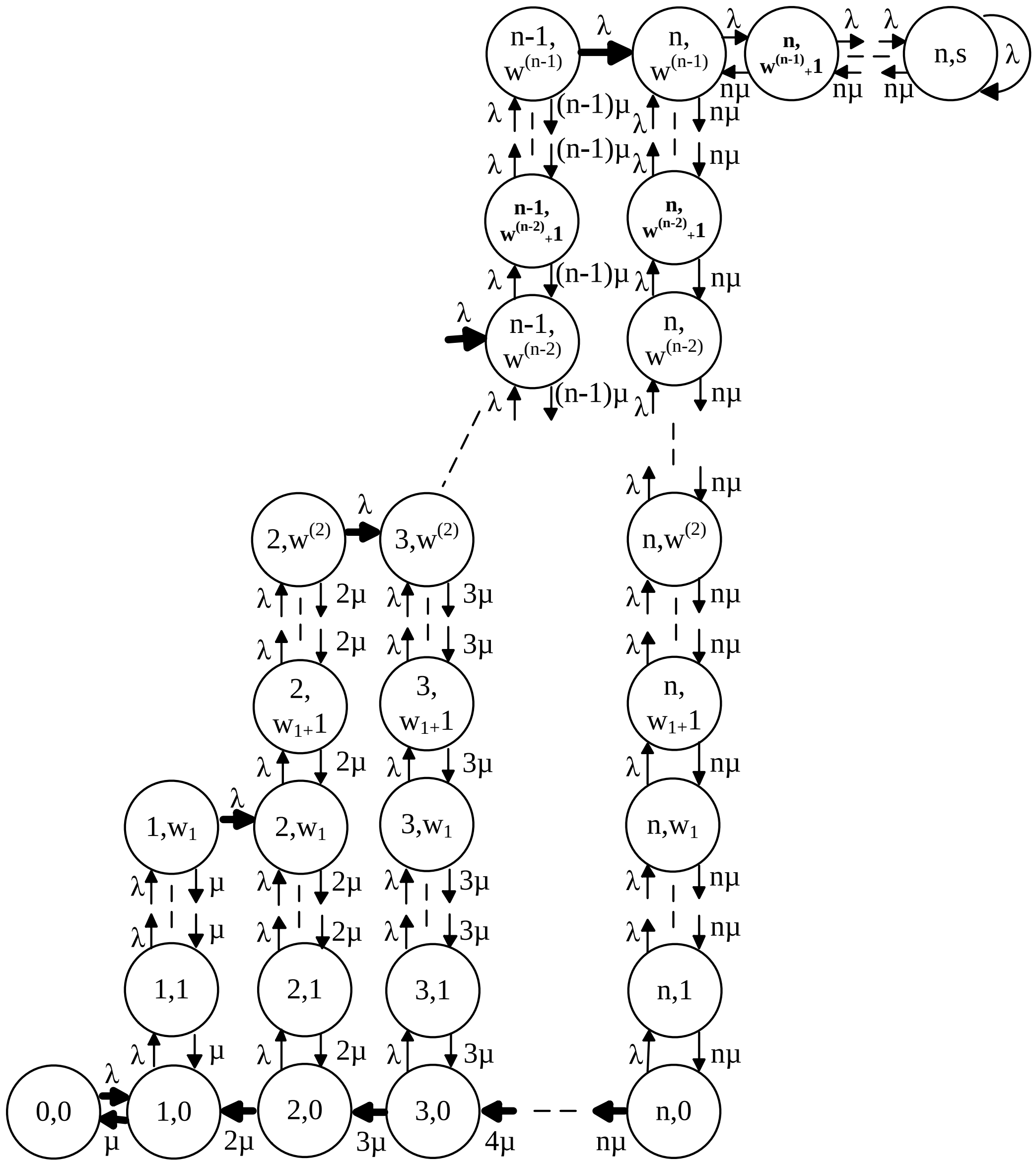
s: maximum buffer capacity

SELF-ADAPTING MODEL BY FSM

(2) MULTIPLE SERIAL HYSTERESIS MODEL



SELF-ADAPTING MODEL BY FSM / (3) MULTIPLE PARALLEL HYSTERESIS MODEL



5. PERFORMANCE ANALYSIS AND RESULTS (1)

MODEL ASSUMPTIONS

- Load-Dependent Activation / Deactivation of Resources -
- Multiple Parallel Hysteresis Model with Server Activation Overhead
- Server Activation: after Server Booting, Queue Threshold Crossing, Process Migration
- Server Deactivation: only when a Server Becomes Idle or the System Becomes Empty (Server Consolidation)
- Notations:

λ	Arrival Rate (Requests, Data Units, ...)
μ	Service Rate of a Server
α	Activation Rate of a Triggered Server Activation
ρ	Utilization Factor ($\rho = \alpha/\mu$)
$E[T_W \mid T_W>0]$	Mean Waiting Times of Delayed Requests
R_A	Server Activation/Deactivation Rate
W(>t)/W	Compl. DF of Buffered Requests

5. PERFORMANCE ANALYSIS AND RESULTS (2)

NUMERICAL EVALUATION

- 1st Choice: Based on the fundamental solutions of Ibe/Keilson by Green’s Function

  - **Result:** Numerically too complex

- 2nd Choice: Based on the fundamental solutions of Lui/Golubchik by Stochastic Complement Analysis

  - **Result:** Numerically too complex

- 3rd Choice: New solution by iterative recursions

  - **Result:** Extremely fast and numerically stable
    Extension to DF of delays
    Optimization wrt performance constraints

- Extensions

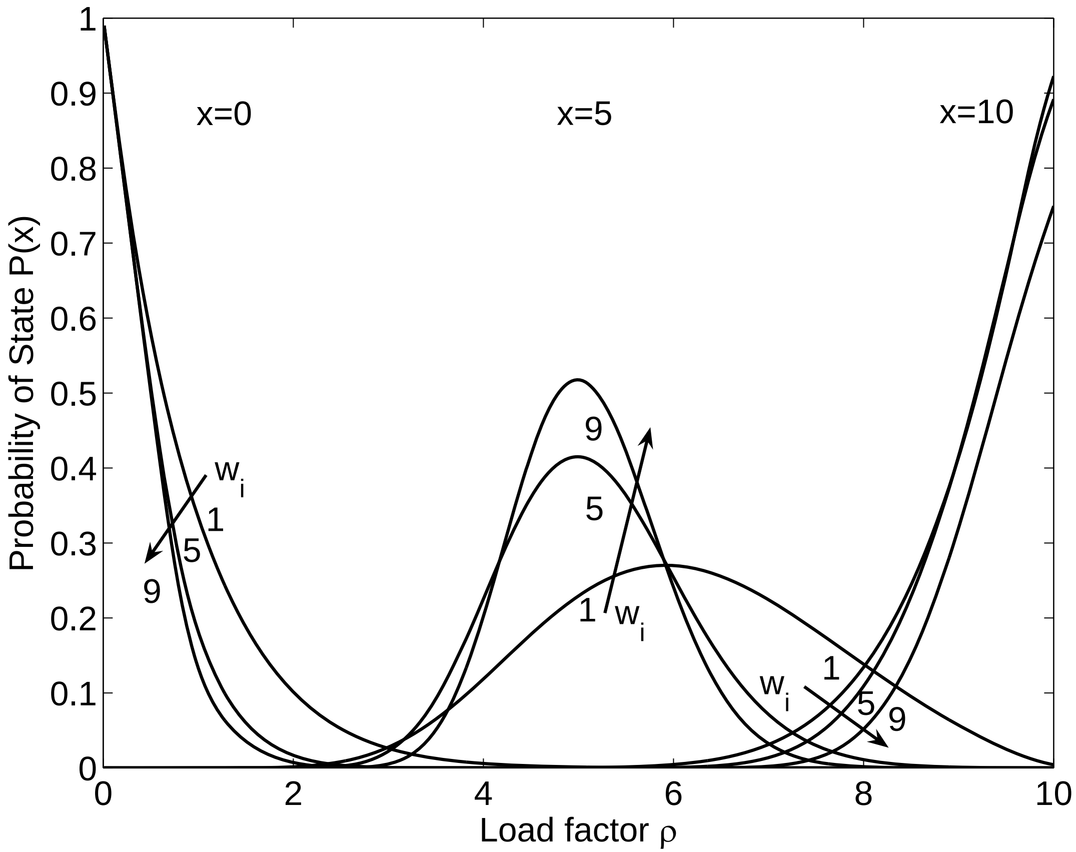
  In all solution methods certain generalizations are possible as

  – bulk arrivals
  – inclusion of activation overhead
  – inclusion of look-ahead activations

5. PERFORMANCE ANALYSIS AND RESULTS (3)

NUMERICAL RESULTS (One DC only)

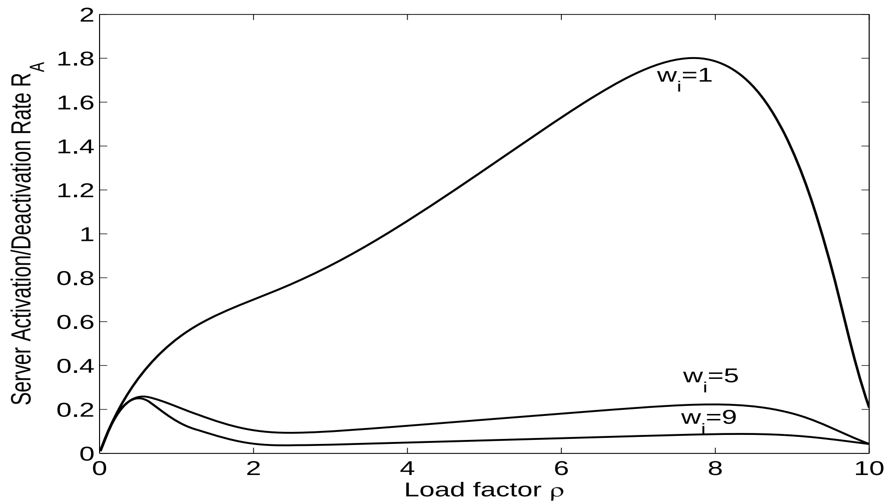
MULTIPLE SERIAL HYSTERESIS MODEL **Probabilities of State**



5. PERFORMANCE ANALYSIS AND RESULTS (4)

NUMERICAL RESULTS (One DC only)

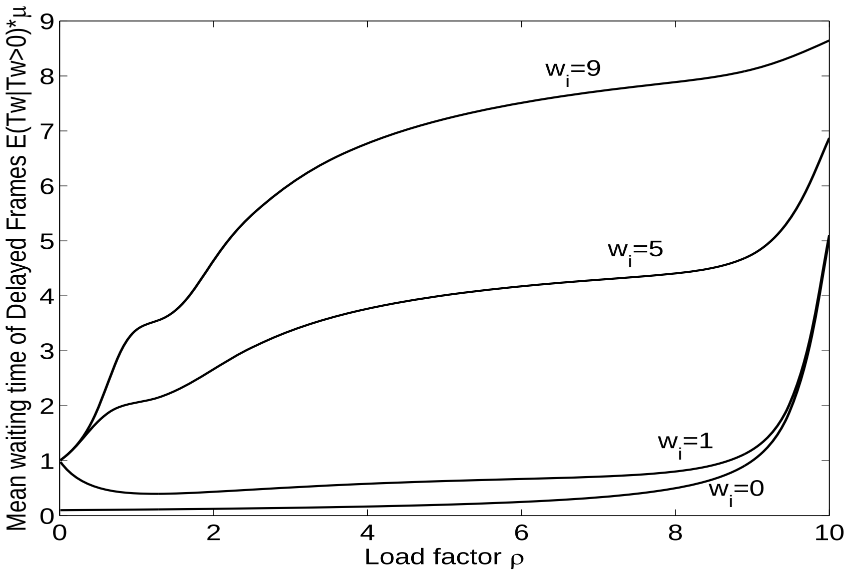
MULTIPLE SERIAL HYSTERESIS MODEL **Server Activation / Deactivation Rate**



5. PERFORMANCE ANALYSIS AND RESULTS (5)

NUMERICAL RESULTS (One DC only)

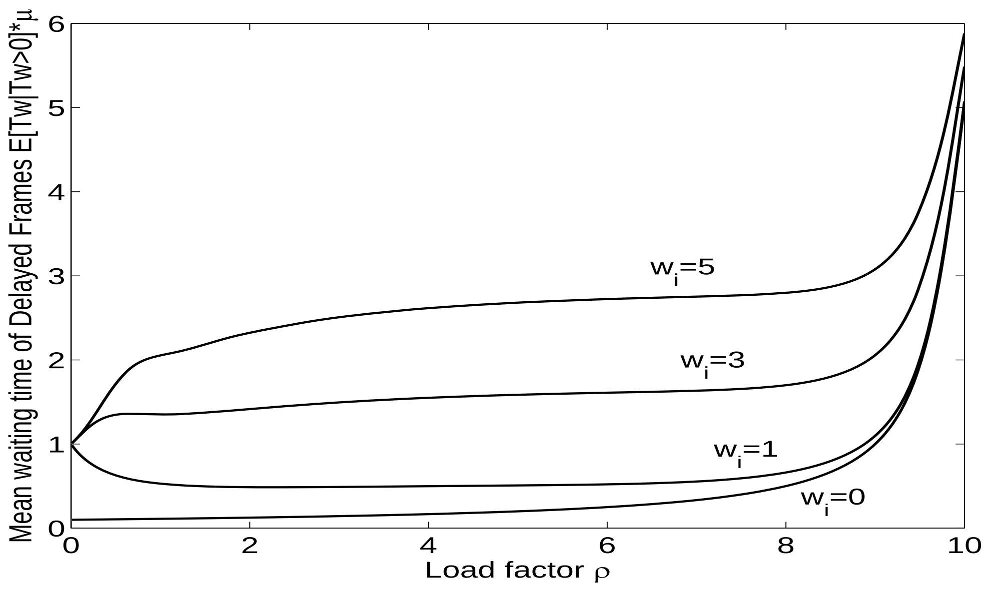
MULTIPLE SERIAL HYSTERESIS MODEL **Mean Waiting Time of Delayed Requests**



5. PERFORMANCE ANALYSIS AND RESULTS (6)

NUMERICAL RESULTS (One DC only)

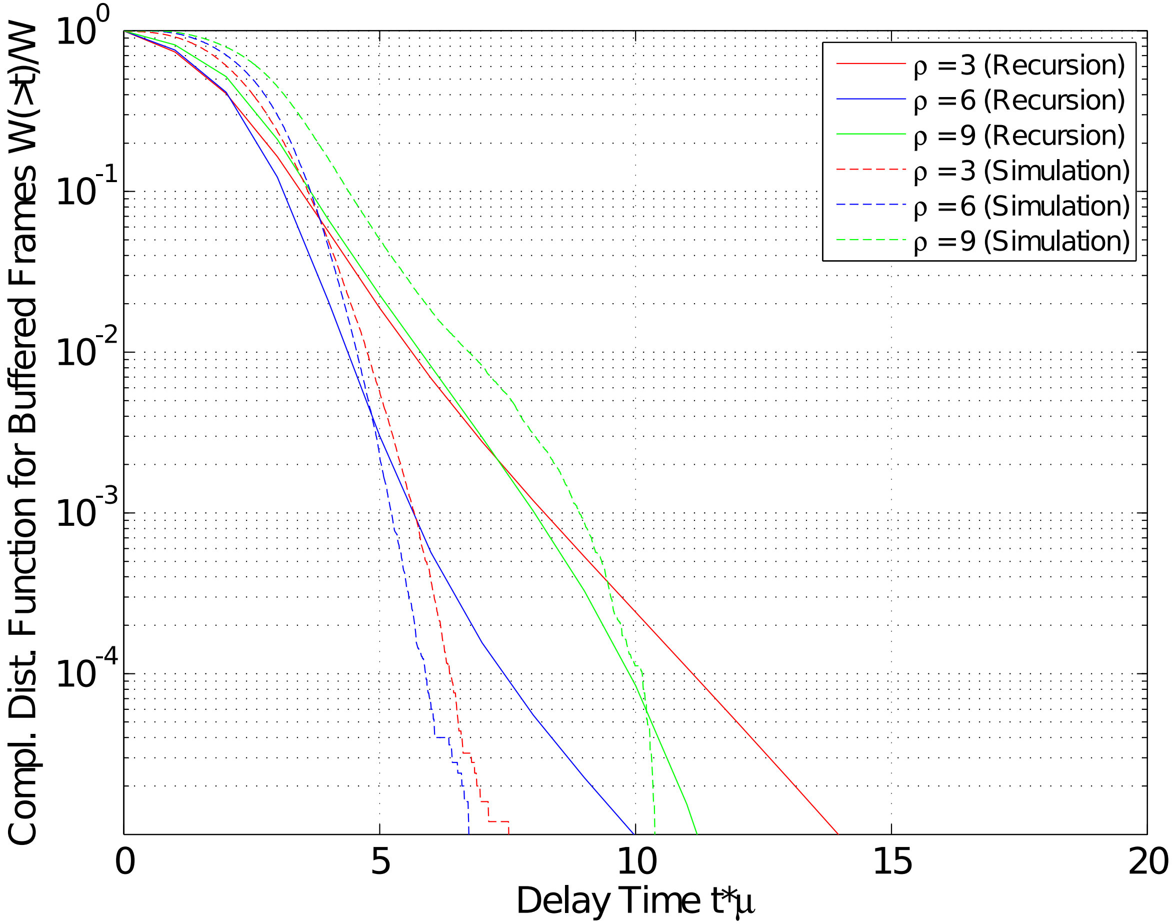
MULTIPLE PARALLEL HYSTERESIS MODEL **Mean Waiting Time of Delayed Requests**



5. PERFORMANCE ANALYSIS AND RESULTS (7)

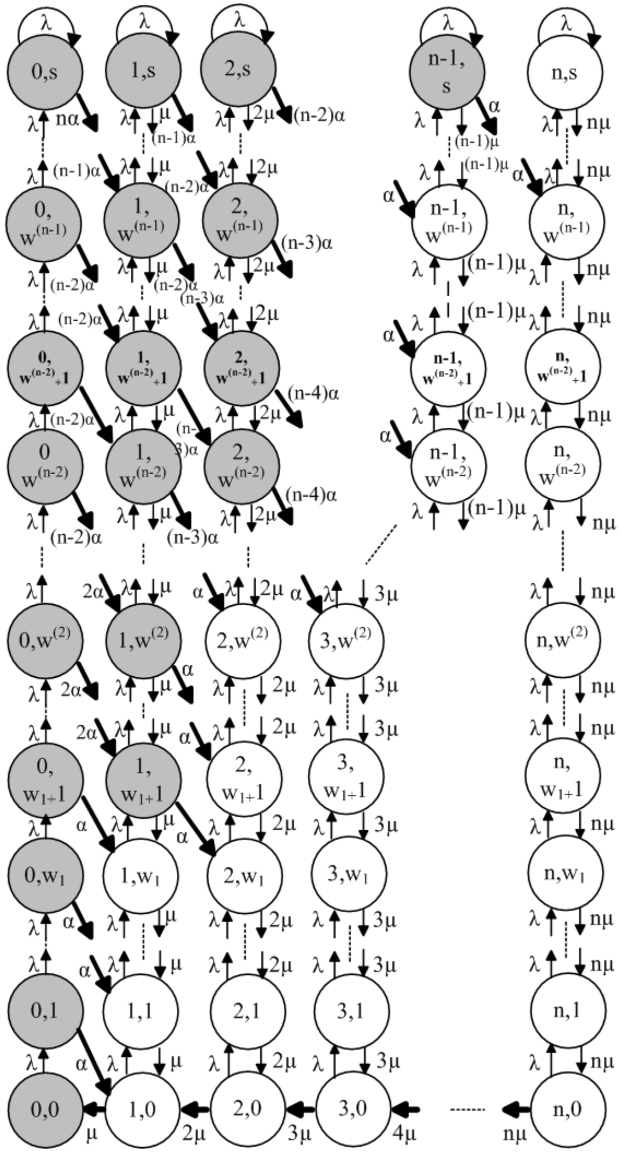
NUMERICAL RESULTS (One DC only)

MULTIPLE PARALLEL HYSTERESIS MODEL **Compl. DF of Buffered Requests**

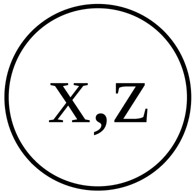


Conditions for the FSM Control:

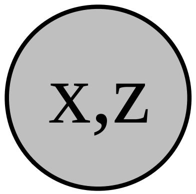
- Multiple hysteresis thresholds for automatic adaptation to variable load
- Buffering of requests to throttle down frequent server activations
- Serving of tasks with maximum possible service rates by activated servers
- Throttling of server deactivations by Dynamic Frequency Scaling (DFS)
- Two server deactivation modes:
  - Server Cold Standby (CSB) ---> Booting required for activation
  - Server Hot Standby (HSB) (Sleeping Mode) ---> Warmup required for activation (Realized by Dyn. Voltage Scaling, DVS)
- All requirements can be met by a pseudo-2-dimensional FSM
- Exact analysis by fast recursive algorithm under Markovian traffic Assumptions
- Parameters:
  - λ task (job) arrival rate ($1/\lambda$ mean interarrival time)
  - μ task service rate ($1/\mu$ mean service time)
  - α server activation rate ($1/\alpha$ mean activation time for booting/warmup)
  - μ^* reduced service rate by DFS



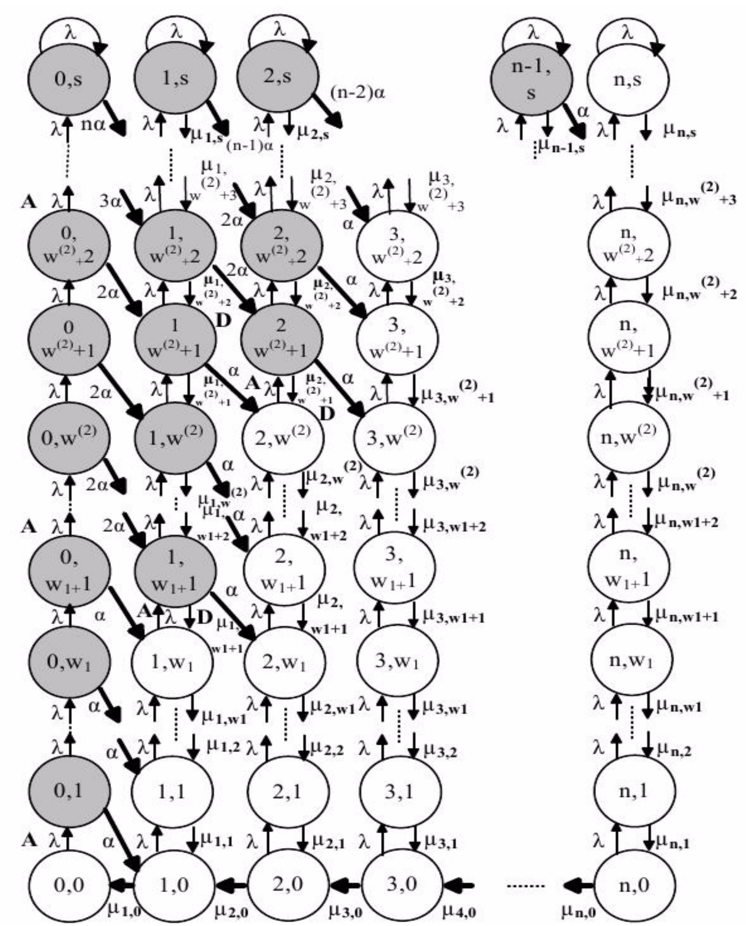
- Multiple Parallel Hystereses Multi-Server Queuing System with/without Activation Overhead



without Activation Overhead



with Activation Overhead



- Multiple Parallel Hystereses Multi-Server Queuing System with/without Activation Overhead and DFS

NUMERICAL RESULTS (One DC only): *Probability State Distributions*

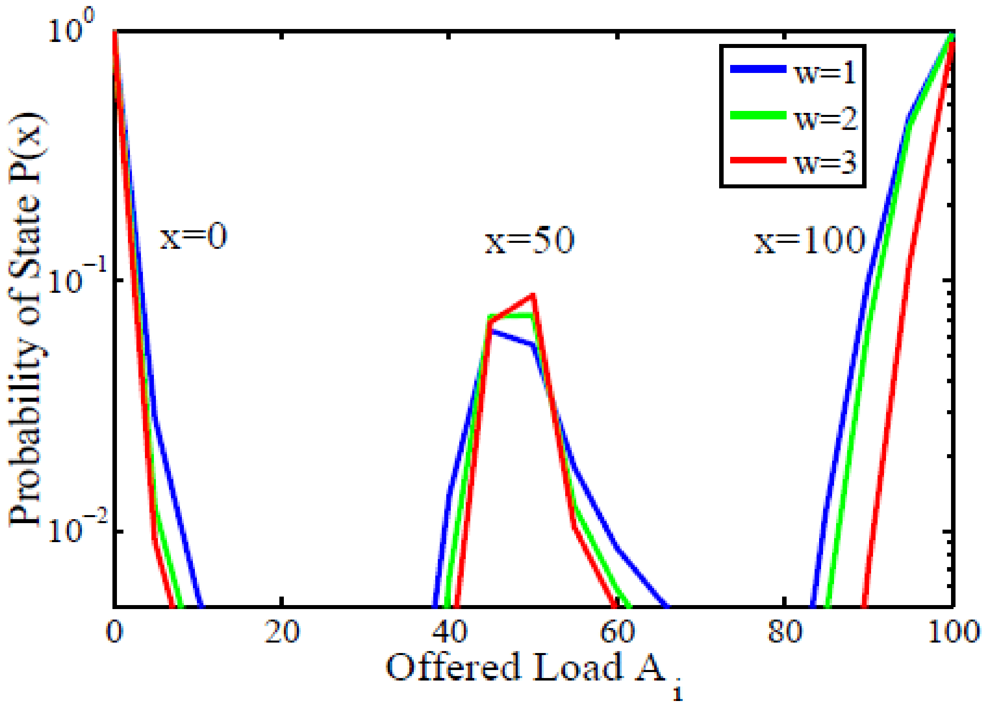


Figure: Probability of 'x' active servers vs. offered load
n = 100, s = 300, α = 1, variable w

NUMERICAL RESULTS (One DC only): *Server Activation Rate*

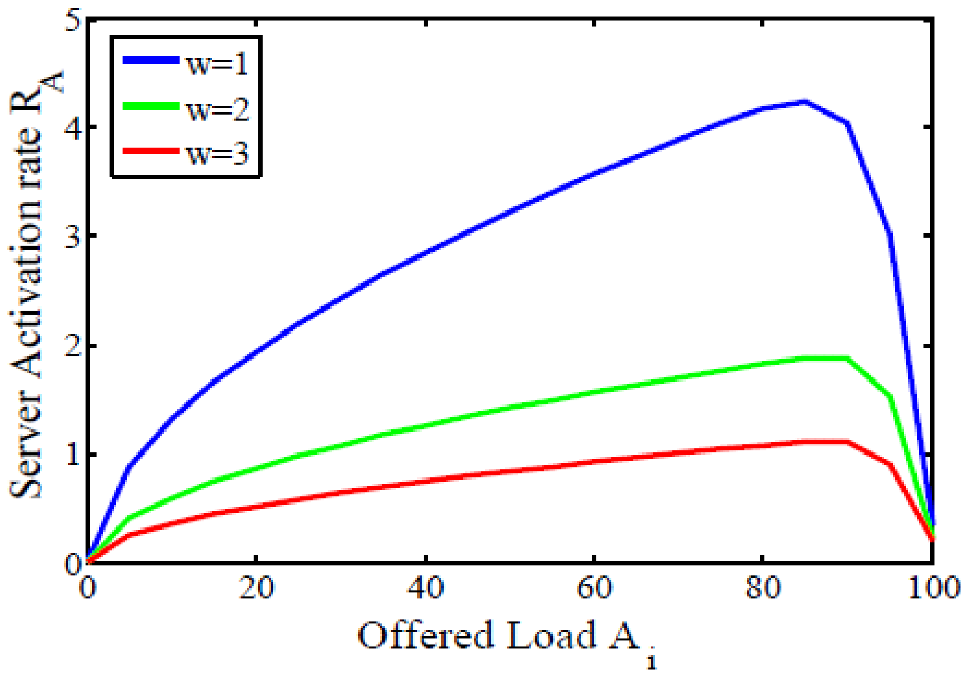


Figure: Server activation rate vs. offered load

$n = 100$, $s = 300$, $\alpha = 1$, variable w

NUMERICAL RESULTS (One DC only): *Mean Delay*

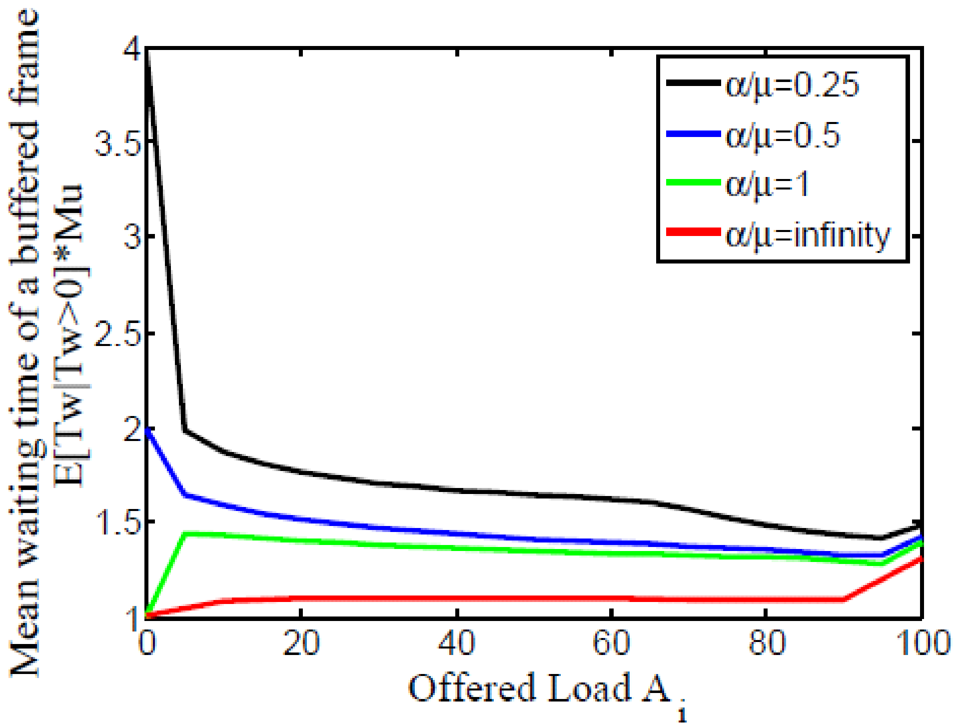


Figure: Mean delay of delayed frames vs. offered load

n = 100, s = 200, w = 2, variable α/μ

STATIC LOAD BALANCING ALGORITHM

- Algorithm Steps

1. Determine the maximum load that could be handled by each data center
   $A_{(max,i)}$ = [function (n_i) | $t_W < t_{SLA}$]
2. Determine the load margin $\Delta A(i) = A_i - A_{(max,i)}$
   If $\Delta A(i) > 0$: Data center i is overloaded and the extra load $\Delta A(i)$ needs to be shifted to another data center.
   If $\Delta A(i) \leq 0$: Data center i can still handle extra load equal to $\Delta A(i)$ without affecting its performance.
3. For DCs whose $\Delta A(i) > 0$, shift this amount of load to the nearest DC who can accommodate this load shift, fully or partially.
4. Repeat the above steps until no more load shifting is necessary.

DYNAMIC LOAD BALANCING ALGORITHM

- Assumptions and Migration Condition
  - N data centers are involved in the load balancing process
  - Each data center has n_i servers and load $A_i = \lambda_i/\mu_i$
  - 2-dimensional FSM, states (x_i,z_i), x_i # of busy servers, z_i # buffered jobs
  - Each data center is operated according to the Multiple Parallel Hystereses
  - Data centers distribute their actual man job waiting times $E[T_{Wi}]$ periodically
  - Time for a process (job) migration to another DC t_m
  - Service level agreement (QoE) by job waiting time threshold t_{W0}
  - Logical condition C job migration (C = TRUE):

$$C = \left(\frac{z_i}{n_i \mu_i} \geq t_{W0}\right) \wedge \left(E[T_{Wj}] + t_m < \frac{z_i}{n_i \mu_i}\right) \text{ for all } j \neq i$$

8. SUMMARY AND OUTLOOK

- Internet Paradigm Shift: Information Transport ----> Information Centric Network
- Cloud Server Virtualization allows for Flexible Content Distribution and Access
- Network Named Content vs. Network Caching
- Models for Self-Adapting DC Server Activation/Deactivation
- Trade-off between Power Reduction and Performance
- Algorithm for Load Balancing and Server Consolidation

Outlook

- Realistic Cloud Application Classes
- Refined Models for DC Architectures and Operations
- Cost Optimization