ZUR ROLLE DER ENERGIEEFFIZIENZFORSCHUNG BEI DER ENERGIEWENDE
- Allgemeine Herausforderungen und Lösungsbeispiele aus der IKT -

Paul J. Kühn
Institut für Kommunikationsnetze und Rechnersysteme (IKR)
Universität Stuttgart
paul.j.kuehn@ikr.uni-stuttgart.de; www.ikr.uni-stuttgart.de

Keynote, Fachtagung SMART ENERGY 2013
Dortmund, 14. - 15. November 2013

ENERGIEWENDE (1) Erneuerbare Energien

- Endlichkeit fossiler Brennstoffe (Kohle, Erdöl, Erdgas)
- Problematik der Nuklearenergie (Tschernobyl, Fukushima)
- Erneuerbare Energien

  - Windparks
  - Photovoltaik
  - Holzvergasung
  - Biomasse

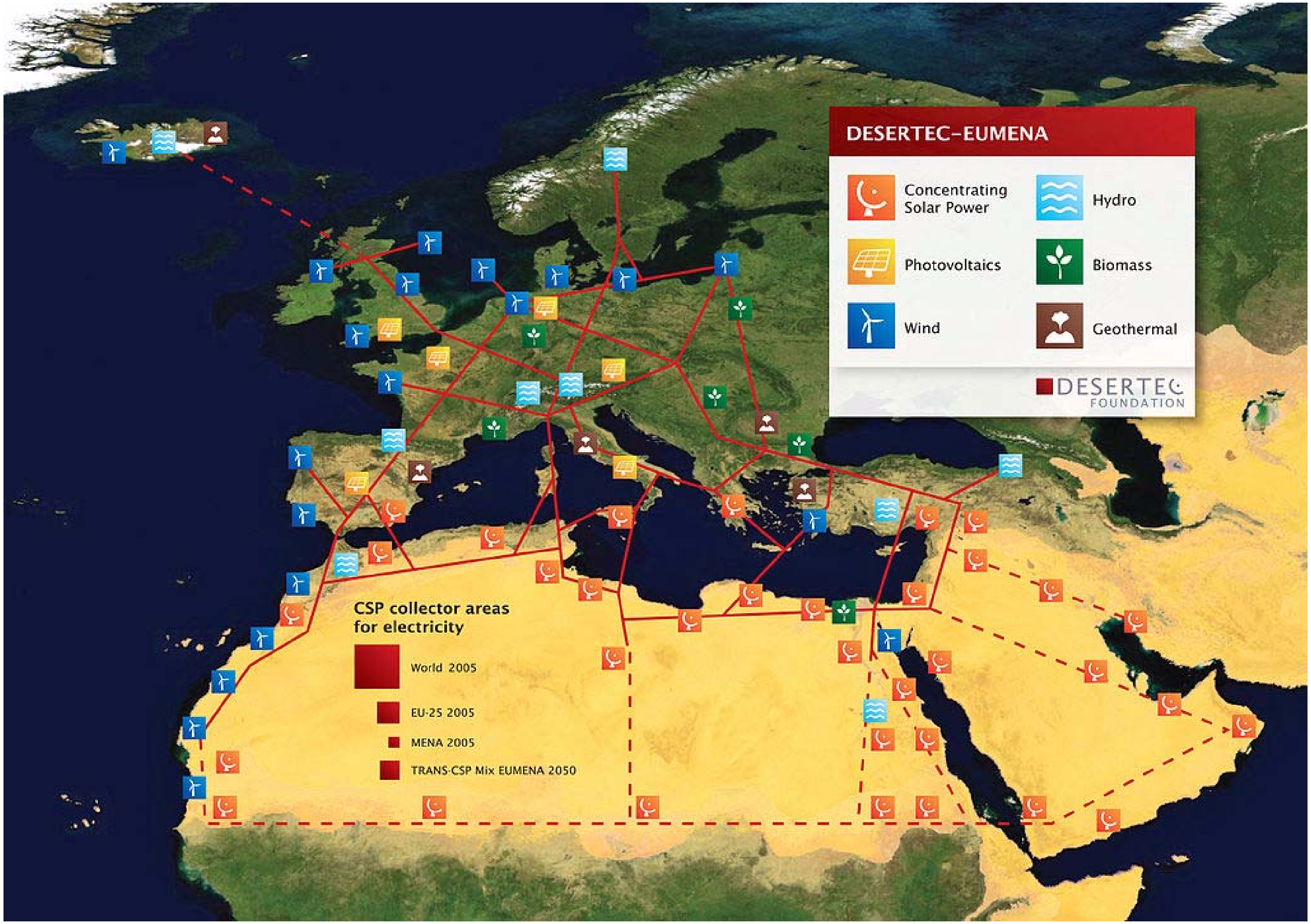
"Fukushima"

"Windparks und Photovoltaik"

"Solarthermie"

"Holzvergasung und Biomasse"

"Netzausbau" Beispiel DESERTEC



ENERGIEWENDE (2) Herausforderungen (1)

- Volatilität erneuerbarer Energien (Windenergie, Photovoltaik/Solarthermie)

- Energiespeicherung
  - Begrenzte Möglichkeit zur Speicherung elektrischer Energie
  - Energiewandlung durch Lageenergie ---> Umfang begrenzt, mittelfristig
    "Pumpspeicherwerke"
  - Energiewandlung in chemische Energie ---> niedriger Wirkungsgrad
    "Power-to-Gas", "Methanisierung" Erdgasnetz als Speicher
    "Brennstoffzellen" ---> mittel- bis langfristig
  - Energiewandlung in Wärme ---> Wirkungsgrad

- Energieübertragungsnetze
  - Anbindung von Off-Shore Windparks ---> Genehmigungsverfahren
  - Hochspannungs-Gleichstromübertragung (HGÜ) ---> Kosten
  - Netzsynchronität Komplexe Regelung

ENERGIEWENDE (2) Herausforderungen (2)

- Energieverteilnetze
  - Bidirektionale Energieflüsse ---> Ausbau der Netzinfrastruktur
  - Messstellenbetrieb Wirtschaftlichkeitsfragen
  - Energiemarkt

- Intelligente Steuerung des Energieverbrauchs
  - Smart Grid ---> Energieinformationsnetze
    Netztechnologien
    Datenerfassung, -speicherung und -verarbeitung
    Netzsicherheit und Datenschutz

- Energieeffizienz ---> Rationelle Verwendung el. Energie

Teil II: ENERGIEEFFIZIENZFORSCHUNG Beispiel: IKT (1)

Hauptverbraucher in der IKT:

- Große Datenzentren (Data Centers, DC)
  - Server-Farmen aus hunderttausenden von Prozessoren
  - Massenspeicher
  - Kühleinrichtungen
- Netzwerkknoten
  - Vermittlungseinrichtungen (Switch)
  - Router
- Mobilfunk-Basisstationen
- Übertragungseinrichtungen
- Endgeräte (Festnetz, Mobilfunknetze, WLAN)

ENERGIEEFFIZIENZFORSCHUNG Beispiel: IKT (2)

Gesamtenergiebedarf zwischen 5 und 10 % des Weltenergiebedarfs!

Wachstum des Internetverkehrs ist exponentiell

Faktor 100 in den nächsten 10 Jahren!

Kann der technologische Fortschritt mit diesem Wachstum Schritt halten?

→ Energieeffizienzforschung ist dringend erforderlich

→ Hohes Forschungsinteresse beobachtbar

- Forschungsprogramme: EU, BMBF
- Industrie-Initiativen: Green Touch
- Int. Fachtagungen: IEEE, ACM, IFIP, ...

→ Nachhaltigkeit über den gesamten Lebenszyklus von Produkten
Rohstoffe, Fertigung, Betrieb, Recycling

--> Gütesiegel werden zum wirtschaftlichen Erfolgsfaktor!

ENERGIEEFFIZIENZFORSCHUNG Beispiel: IKT (3)

Beispiel 1 Lastadaptive Aktivierung/Deaktivierung von Cloud DC-Servern

PANEL Int. Teletraffic Congress (ITC 25)

Shanghai/China, September 2013

Beispiel 2 Endgeräte mit Schlafmodi und dynamischer Aktivierung

Paper Int. Conf. on Communications and Electronics (ICCE)

Hue/Vietnam, August 2012

THE USE OF PERFORMANCE MODELING METHODS FOR ENERGY EFFICIENCY AND QoS MANAGEMENT OF ICT SYSTEMS AND IN SMART GRID CONTROL

Paul J. Kuehn
Institute of Communication Networks and Computer Engineering (IKR)
University of Stuttgart, Germany
paul.j.kuehn@ikr.uni-stuttgart.de; www.ikr.uni-stuttgart.de

Keynote
24th Tyrrhenian International Workshop 2013 on
Digital Communications: Green ICT
(24th TIWDC 2013), Genova, Italy, September 23-25, 2013

OUTLINE

Part 1: Key Examples of Complex ICT Systems

1.1 Information-Centric Networking

1.2 Smart Grid

Part 2: Modeling Methodology in Complex System Management

Example:

Automatic Energy Efficiency Management of Data Center Servers

1.1-1 INFORMATION-CENTRIC NETWORKING

- Major Application Shifts in the Internet

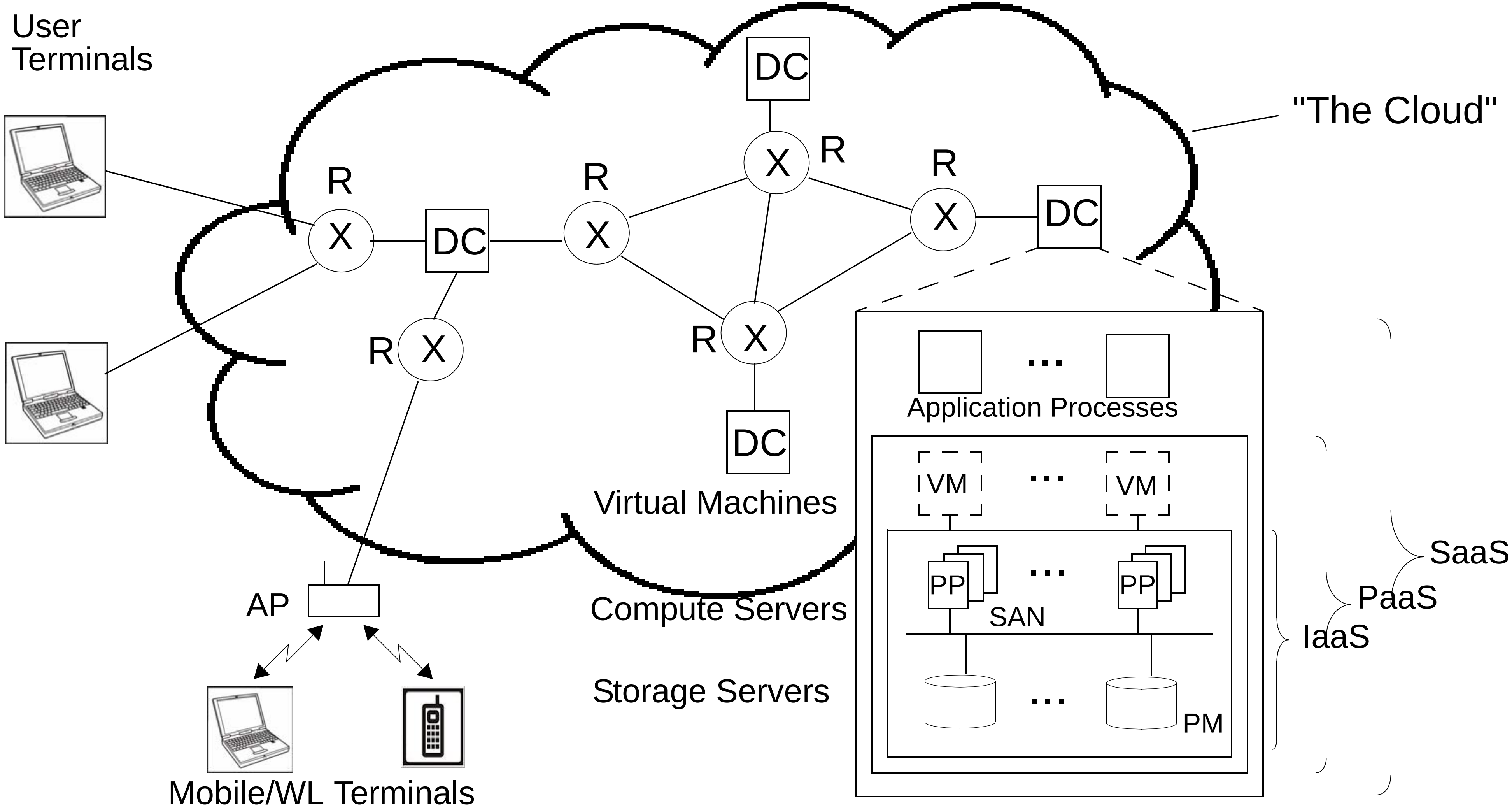
Remote Computation
File Transfer
E-Mail
Content Access (Web)
Peer-to-Peer File Sharing
Voice over IP / Multimedia
Web-based e-Business
Social Networking
Content Access & Distribution

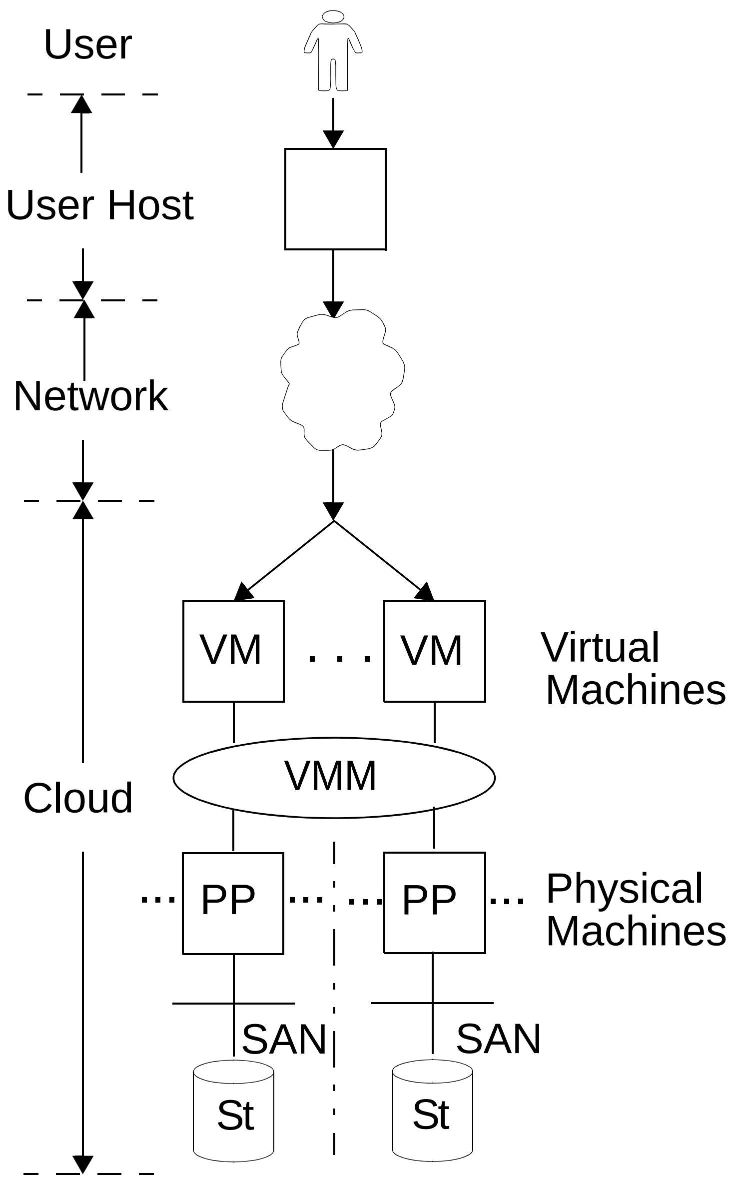
- Paradigm Shifts

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

1.1-2 CLOUD ARCHITECTURES



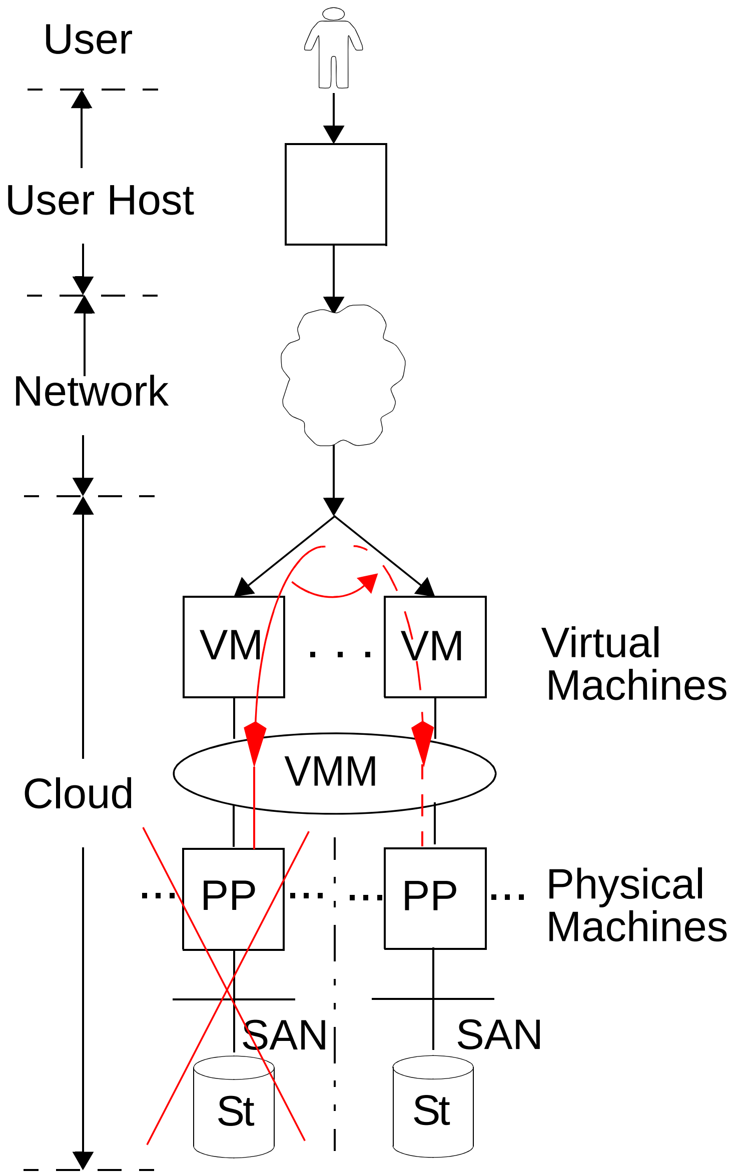


Cloud: Pool of Physical Resources
Interconnected by Network

VM: Virtual Machine
Virtualized View on the Resource Pool

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

1.1-4 MANAGING CDNs - VIRTUALIZATION AND VM MIGRATION



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"

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

1.1-6 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

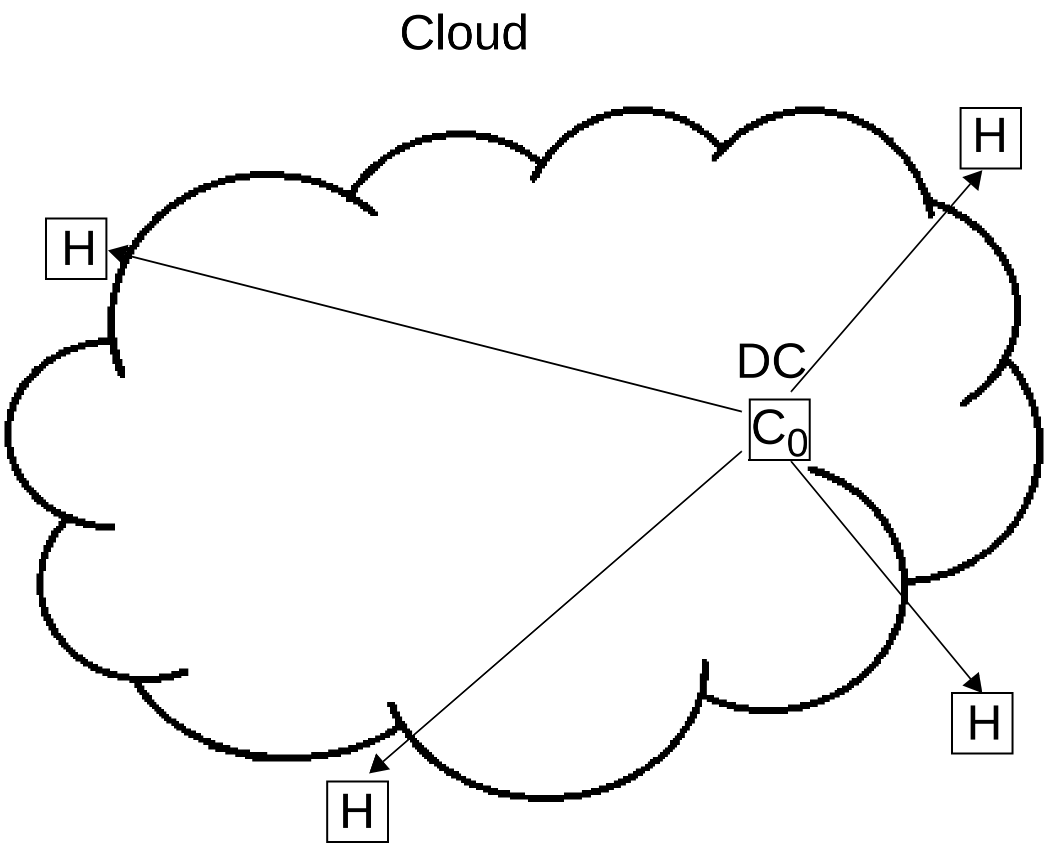
1.1-7 MANAGING CDNs - DYNAMIC PROVISIONING OF PHYSICAL RESOURCES

• 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!)

1.1-8 MANAGING CDNs - CENTRALIZED STORAGE

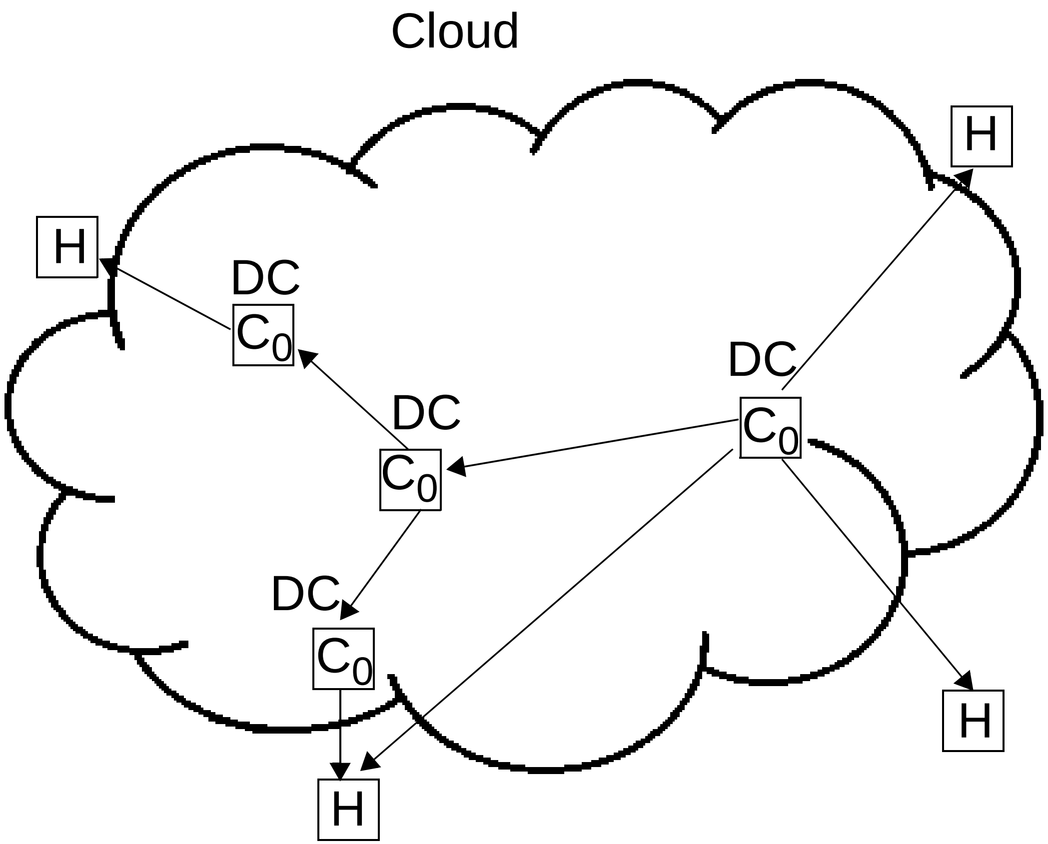


- Multicast Tree
- Minimum Storage Cost
- Maximum Communication Cost
- Maximum Latency
- High Risk, Reliability

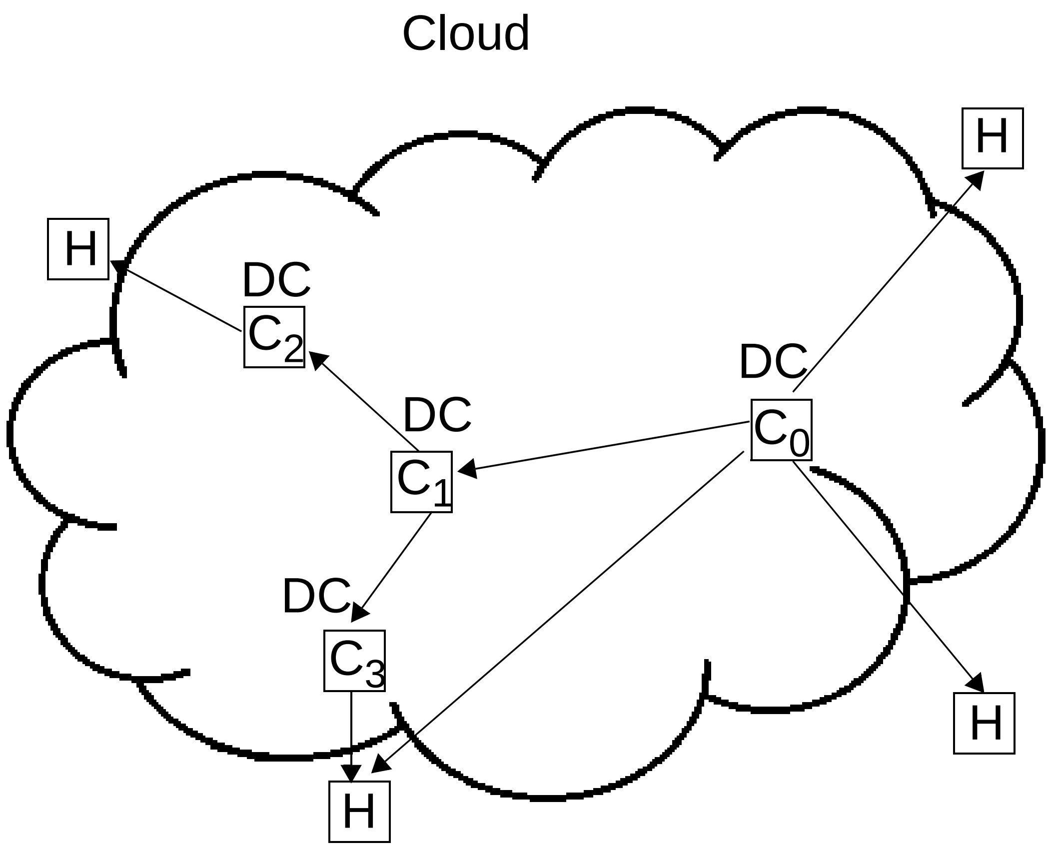
H User Host

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?

1.2-1 Smart Grid - Problem Statement

- Continuous Energy Demand

- Exhaustion of Traditional Energy Resources (Fossile Combustibles)

- Energy Resources
  - limited availability (wind, sun, water, biomass)
  - have to be used jointly

- Reduction of Energy Demand
  - by improved technologies
  - by adaption between energy production and energy usage

1.2-2 Smart Grid - Problem Statement

- Continuous Energy Demand

- Exhaustion of Traditional Energy Resources (Fossile Combustibles)

- Energy Resources
  - limited availability (wind, sun, water, biomass)
  - have to be used jointly

- Reduction of Energy Demand
  - by improved technologies
  - by adaption between energy production and energy usage

Intelligent Coupling of
Energy Production / Conversion and Energy Usage

1.2-3 Smart Grid - Problem Statement

- Renewable Energies
  - Wind Energy ... fast deployment
  - Photovoltaic Energy ... fast deployment
  - Solar-Thermic Energy ... under test/planning (Project Desertech)
  - Biomass Energy ... fast deployment
  - Geothermic Energy ... local applications, big applications questionable
- Energy Storage
  - Pump Water Power Stations ... effective for peak consumption, but limited possibilities; new approaches under research
  - Hydro-Electric Power Stations ... steady supply, economic, but limited possibilities
  - Electric Batteries ... expensive, limited lifetime (number of charging cycles)
  - Methanization ... electric-chemical conversion (water --> H,O; CO_2 --> methan); storage within existing gas network

1.2-4 Smart Grid - Problem Statement

• Current Control: SCADA (Supervisory Control and Data Acquisition and Control)	... limited, too slow
• Standards Development	... > 20 Standards Development Organizations (SDO) NIST, ISO, IEEE, CEN, CENELEC, ETSI, SGIP, ITU-T, IETF, ISGAN, ...
• Measurement Data Acquisition	... huge number of power supply "stations" ... huge number of sensors, measurement data ... coherent real-time data problems GPS-based clocks and location information Phasor-based precise measurements
• Control	... Distributed Wide-Area Control System (WAMS) frequency and voltage control criticality
• Power Management	... by use of stored energy, by intelligent consumption scheduling, or both ... real-time operation

References: Special Issue IEEE Proceedings, June 2011, Vol. 99, No. 6 "Smart Grid"

1.2-5 Smart Grid - Problem Statement

• Smart Meter Technology	... consumption monitoring ... energy consumption data (which, when, preciseness, ...?)
• Control	... remote control (based on energy availability) ... local control (based on tarif or policy?) ... local batteries? ... seasonal experiences
• Field Applications	... manifold in USA, Germany, ...

References: Actual Conference Series

1., 2. and 3. ACM Conference e-energy 2010, 2011, 2012, 2013
1., 2. and 3. Conference Smart Energy, Dortmund, 2010, 2011, 2012, 2013
VDE Congress 2012, Stuttgart
IEEE ICC/GLOBECOM Workshops 2009, ff
... and many others

1.2-6 Smart Grid - Problem Statement

• Electric Transport Networks	... HV AC/DC transmission networks local distribution networks energy storage and energy conversion
• Communication Networks	... existing telecommunication networks power line communications (PLC) wireless communications
• Protocols	... standards by NIST, CENELEC, IEEE, ...
• Data Centers	... monitoring, storage and processing outage analysis and flexible reconfiguration cloud computing
• Smart Home	... home/factory automation temperature control, air conditioning presence detetection control of energy consumption
• Smart City	... traffic control public transport systems car-car, car-infrastructure comms. electro-mobility

1.2-7 Smart Grid - Problem Statement

• Security of Data	... safe transmission safe processing access control
• Risk Analysis	... cyber attack detection and protection man-in-the-middle attacks redundant infrastructures/resilience
• Privacy Problem	... legal requirements ("informationelle Selbstbestimmung") user protection access/use of user energy consumption data profiling of users/user groups/enterprises

References: Claudia Eckert, Christoph Kraus
"Sicherheit im Smart Grid" - Sicherheitsarchitekturen für die Domänen Privatkunde und Verteilnetz unter Berücksichtigung der Elektromobilität
AI Stiftung für Kommunikationsforschung, Nr. 96, 2012

1.2-8 Smart Grid - Problem Statement

- Energy Efficiency ... power management policies
  improved technologies ("Kraft-Wärme-Kopplung")
  low-power device technologies
  data center cooling techniques/placements
  location choices for renewable energies
- Energy-Demand Reduction ... reduction of peak power demands
  elasticity of consume
  intelligent/adaptive power reduction
  (sleep modes, virtualization of cloud servers, ...)
- Optimization and Pricing ... intelligent schedules for energy consumption
  backup energy resource management
  dynamic pricing for energy consumption
  market/auction models (energy as spot market)
  game-theoretic approaches
  . . .

References: Special Issue IEEE Journal on Selected Areas in Communications (JSAC)
July 2012, Vol. 30, No. 6 "Smart Grid Communications"

1.2-9 Smart Grid - Problem Statement

SMART GRID is

- an intelligent energy network and control system of
  - intelligent generators
  - storage facilities
  - loads and transportation equipment
- with the aid of
  - information and communication technologies (ICT) as well as
  - automation technology
- to ensure sustainable and environmentally sound energy supply by means of
  - transparent
  - energy- and cost-efficient
  - safe and reliable

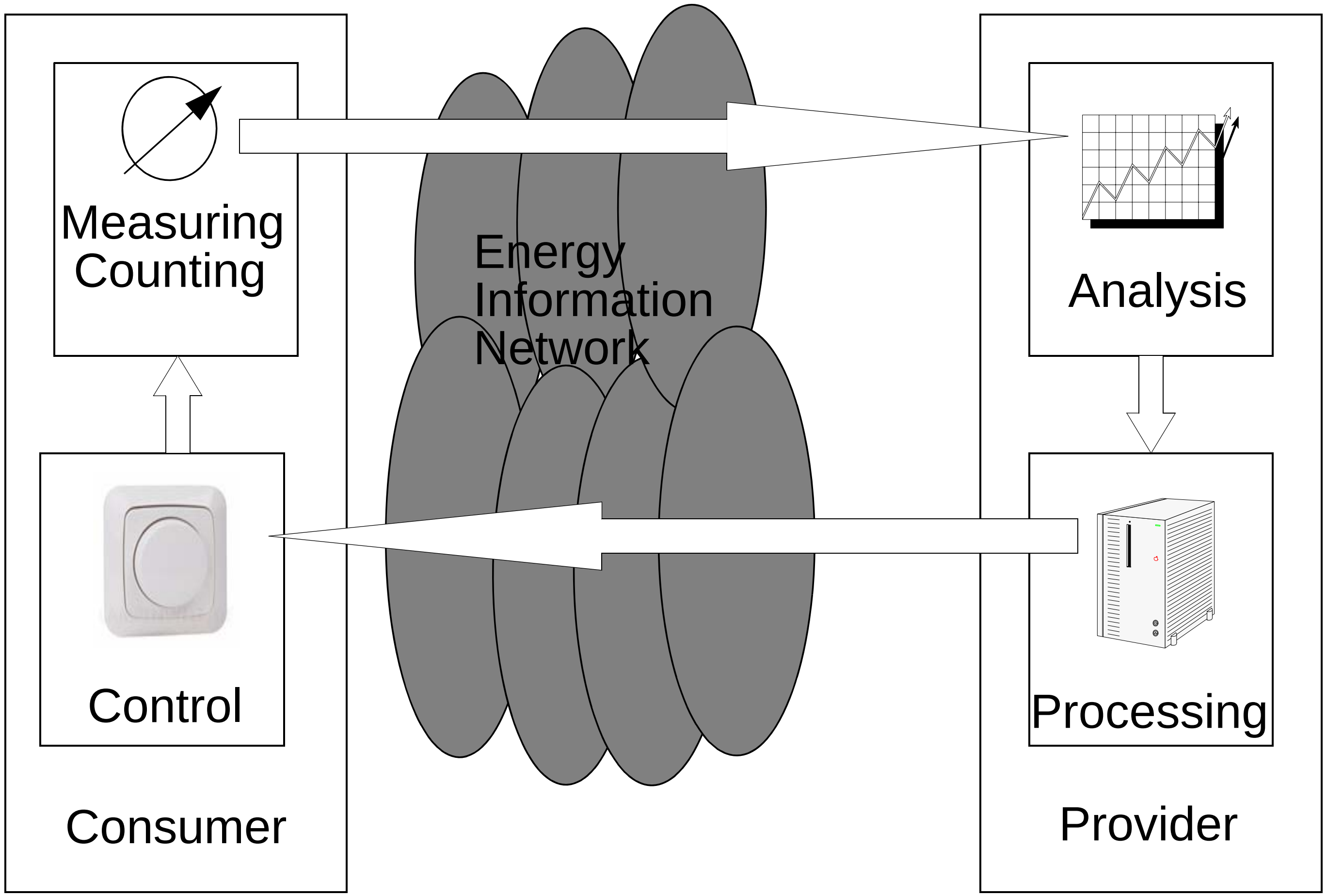
  system operation

1.2-10 Smart Grid - Problem Statement

- Smart Grid is Necessary for the Control of Highly Variable Energy Sources and Energy Consumptions
- Smart Grid is a Technical Challenge
  Cooperation of Energy Sources and Consumers,
  Network Control, Communication Infrastructures,
  Information Processing, Self-Organization, Management, ...
- Smart Grid is a Risk Challenge
  Outages through Breakdowns
  Catastrophic Events
  Cyber Attacks
  Profiling of User Behaviour
  Legal Restrictions

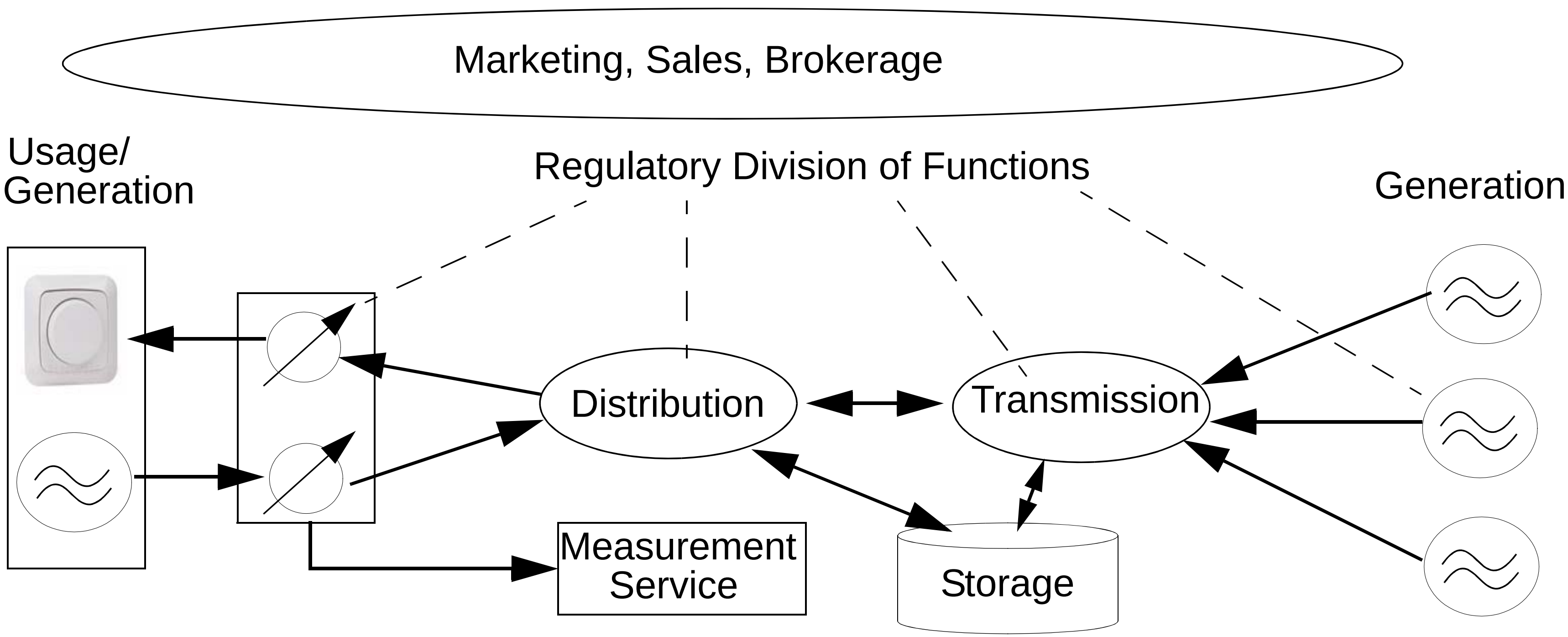
1.2-11 Smart Grid - Problem Statement

Energy Information Network between Producer and Consumer



1.2-12 Smart Grid - Problem Statement

New Role Model



1.2-13 Smart Grid - Problem Statement

Requirements for an Energy Information Network

- Multiple Interfaces (Sensors/Actors, Network Technologies and Protocols)
- Efficiency (Throughput, Response Time)
- Reliability (Breakdown Protection, Resilience)
- Network Management (Konfigurability, Outage Management)
- Distributedness (Decentral/Central Control, Self-Organization, Distributed Software)
- Controllability (Multi-Parameter Control System)
- Context Sensitivity (Detection of Situation by Sensors, Sensor Fusion)
- Security and Privacy (Access Rights, Protection of Personal Data)
- Energy Efficiency (Sensors/Actors, Power Supply)

1.2-14 Smart Grid - Problem Statement

Options

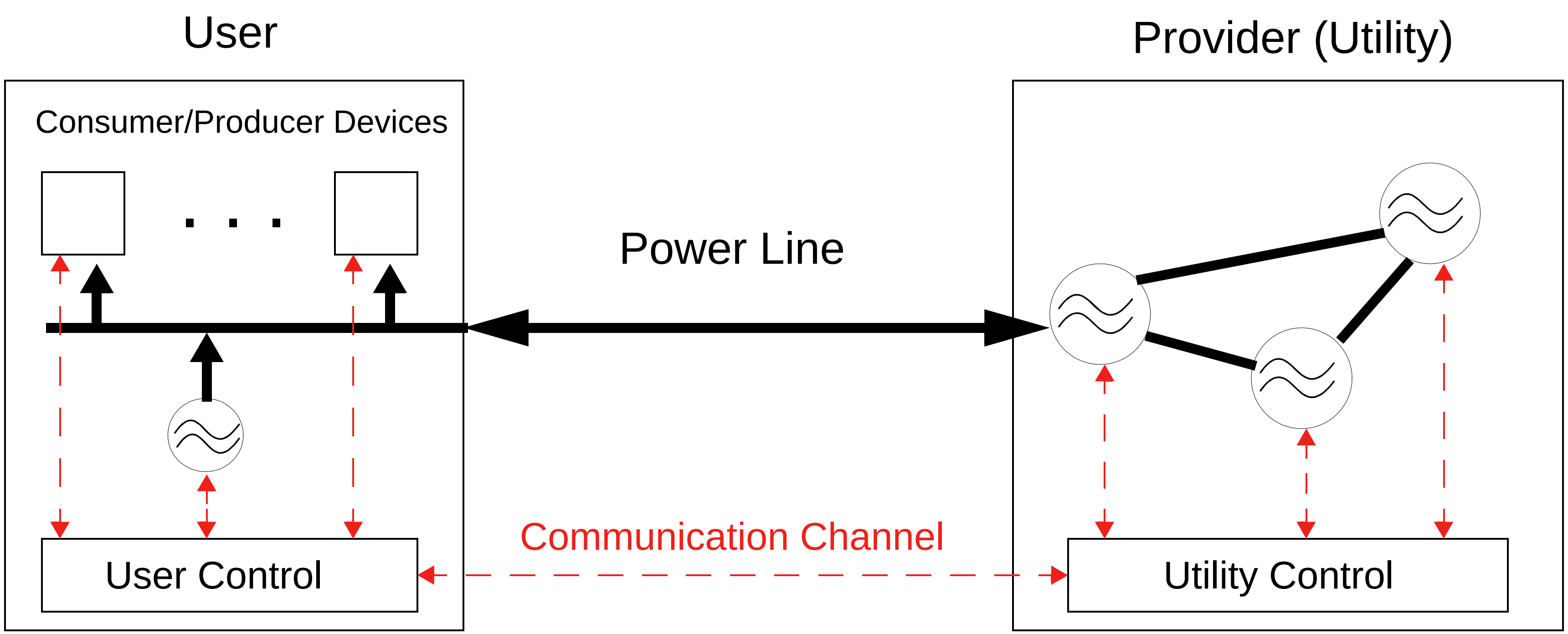
- Inhouse Networks
  - Powerline Communication (PLC)
  - Wireless LANs

  combined with
  - Integrated Telecommunication and Data Networks

- Access Networks and Wide Area Networks
  - Existing DSL Technologies
  - Future Optical Access Networks
  - Metro / Core Networks
  - Fiber Networks of Energy Providers

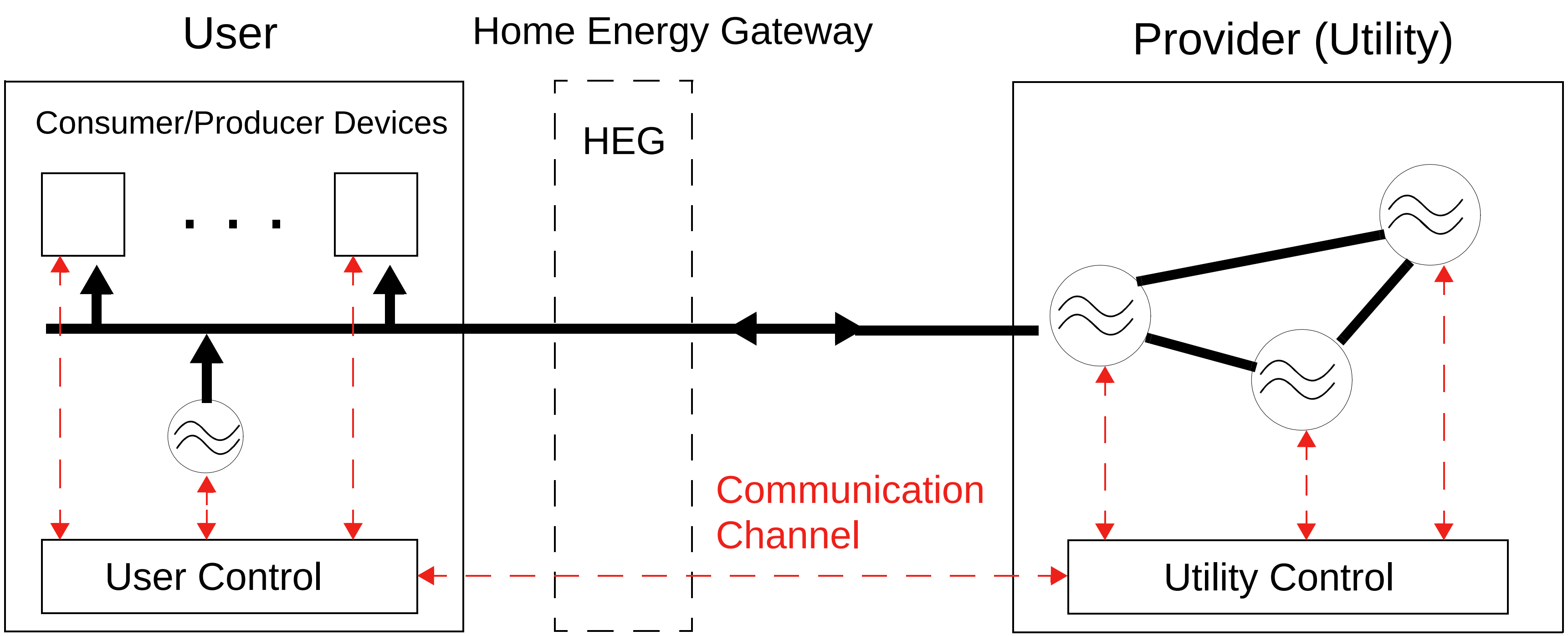
  with Gateways to the Internet

1.2-15 Smart Grid - Problem Statement



- **Ideal Control** - Transparency of En. Consumption ---> Power Requirement P(t)
  - Transparency of En. Production ---> Power Availability (kW, Tariff, ...)
- **Implementation** - Central Control (Provider)
  - Decentral Control (Consumer)
  - Distributed Control (Cooperation between Provider and Consumer)

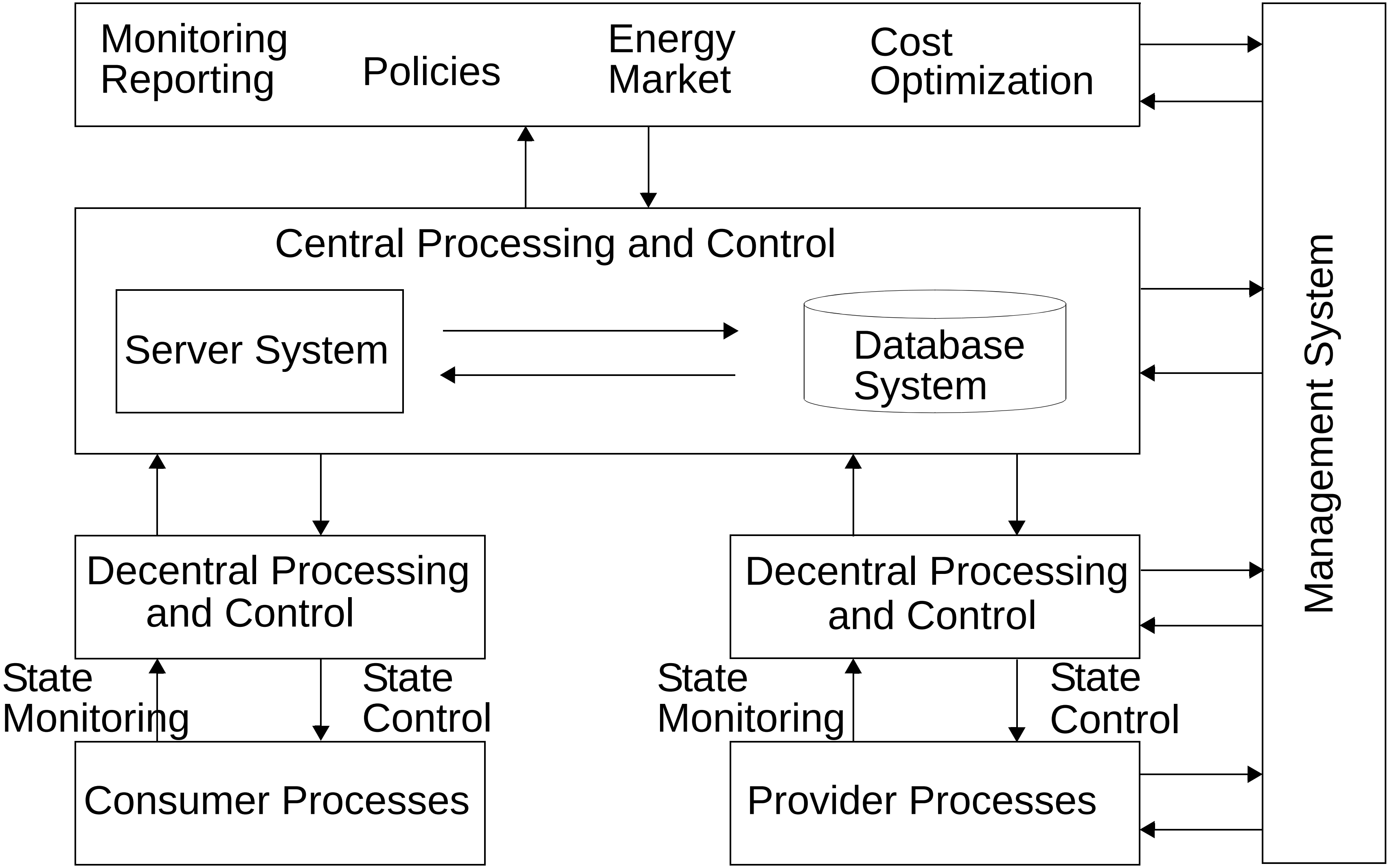
1.2-16 Smart Grid - Problem Statement



HEG Functions Smart Energy Platform + Home Automation

- collects meter data, appliance data, load data, tariff data
- control of home appliances (automation platform)
- security protocols for trusted applications
- user privacy protection

1.2-17 Smart Grid - Problem Statement



1.2-18 Smart Grid - Problem Statement

Smart Grid:		Electrical Energy Network + Communication Overlay Network for Usage Data Acquisition, Processing and Control
Potential:		Optimum Cooperation between Energy Generation and Consumption
Risks:	-	Outage of Communication, Processing and Control by Breakdown of Components
	-	Cyber Attacks through the Communication Network (Cyber Terrorists)
	-	Vulnerability, Threats, Safety, Investment, Inventory
Effected Resources:	-	Processing and Storing Devices
	-	Communication System
	-	Control Systems SCADA-Problem: Supervisory Control And Data Acquisition

1.2-19 Smart Grid - Problem Statement

- **Detailed User Power Consumption Data Underly OECD Privacy Principles**
  - Recording of Smart Meter Data $\Longrightarrow$ Conclusions on Personal/Company Activities
  - Storage Places and Durations of such Data?
  - Access Rights and Control of Usage of such Data?
  - Ownership of such Data?

- **Accuracy, Quality and Control**
  - Accuracy of Measurements or Future Needs
  - Granularity of Energy Consumption Data (Resolution)
  - Completeness and Relevance of Collected Data
  - Application Control (User or Provider?)

1.2-20 Smart Grid - Problem Statement

- **Risk Analysis**
  Device Level, System Level, Network Level
  Novel Mechanisms to Reduce Risks

- **Standardization**
  Requirements (e.g. NIST Standards)
  Protocols and Interfaces

- **Technical Solutions**
  Research Issues
  Development Options
  Prototyping and Field Experiments
  Technical Management

- **Roadmap**
  Strategic Aims
  Evolution Steps

1.2-21 Smart Grid - Problem Statement

- **Sensoric Technology**
  - Smart Meters and Measurement Techniques
  - Home Automation

- **Communication Networks**
  - Cooperation of Sensoric Networks and Communication Infrastructures
  - Network Resilience
  - Cryptographic Protection of User and Control Data
  - Firewall Techniques
  - Identity Management, Virtual Identities
  - Trusted Entities

- **Data Storage and Processing**
  - Management of Access Rights
  - Protection against Physical Damage/Breakdown
  - Protection against Cyber Attacks
  - Secure Cloud Computing and Peer-to-Peer Communication
  - Identity Management
  - Organizational Framework

1.2-22 Smart Grid - Problem Statement

- **Need for Legal Framework**
  - Documentation of Information Types of Collected Data
  - Providing User Notification or User Access to Recorded User Data
  - Limiting the Collected Data to the Required Minimum
  - Limiting the Distribution of Private Consumer Data
  - Limiting the Storage of User Data to the Absolutely Necessary Duration

- **Technical Solutions**
  - Safeguarding of Personal Information Against Unauthorized Access, Disclosure or Copying
  - De-Identifying : Anonymizing / Pseudonymizing of Identities
  - Decentralized Control:

    Smart Power Control Can be Done Completely Decentralized

    Incentives: Online Information about Available Power and Time-Dependent Tariffs

2.1-1 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)

2.1-2 Modeling Algorithms

BASIC IDEAS:

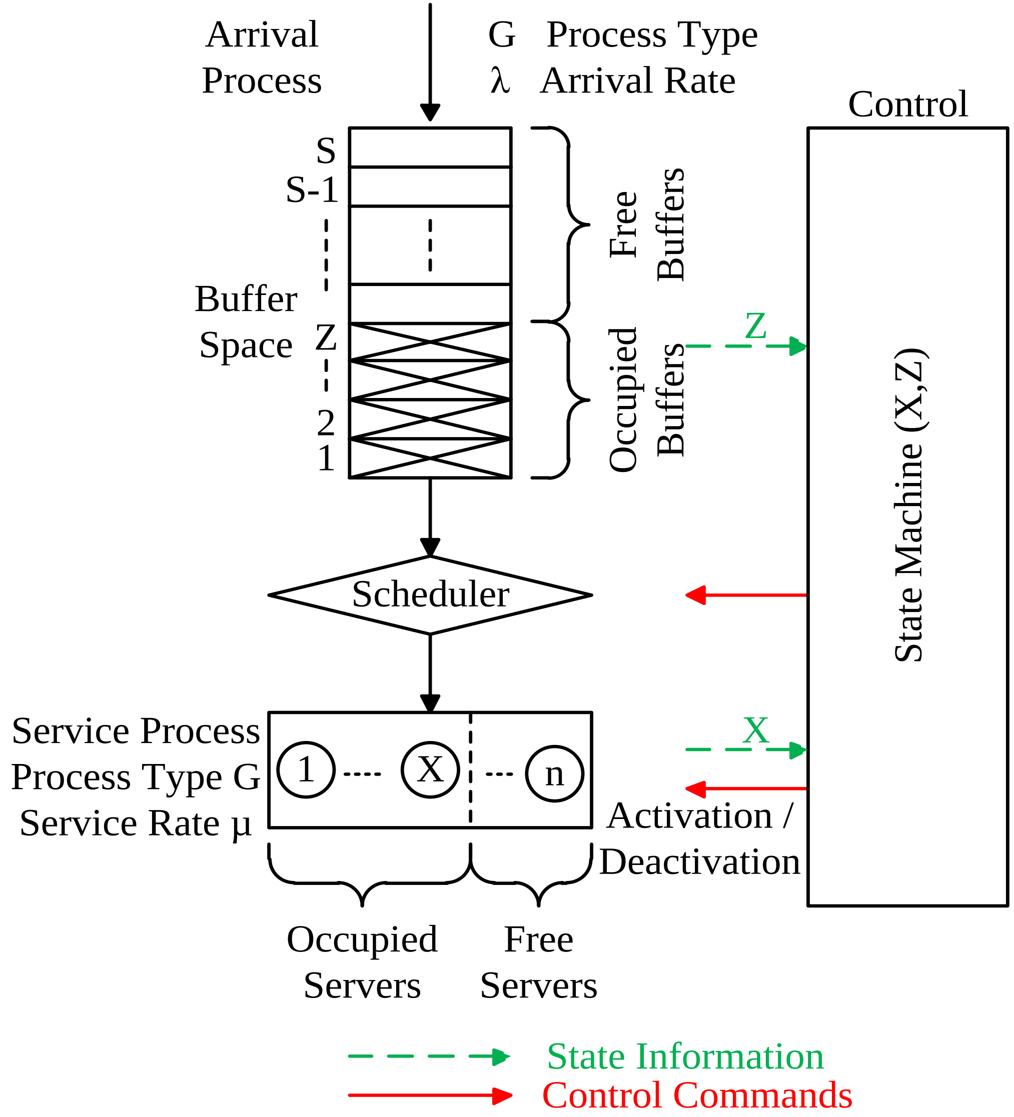
- Self-Adapting Operation of Data Center Resources
  - Local Monitoring of Load Development
  - Local Control of Resource Activation/Deactivation by FSM

BASIC MODEL:

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

2.1-3 Modeling Algorithms

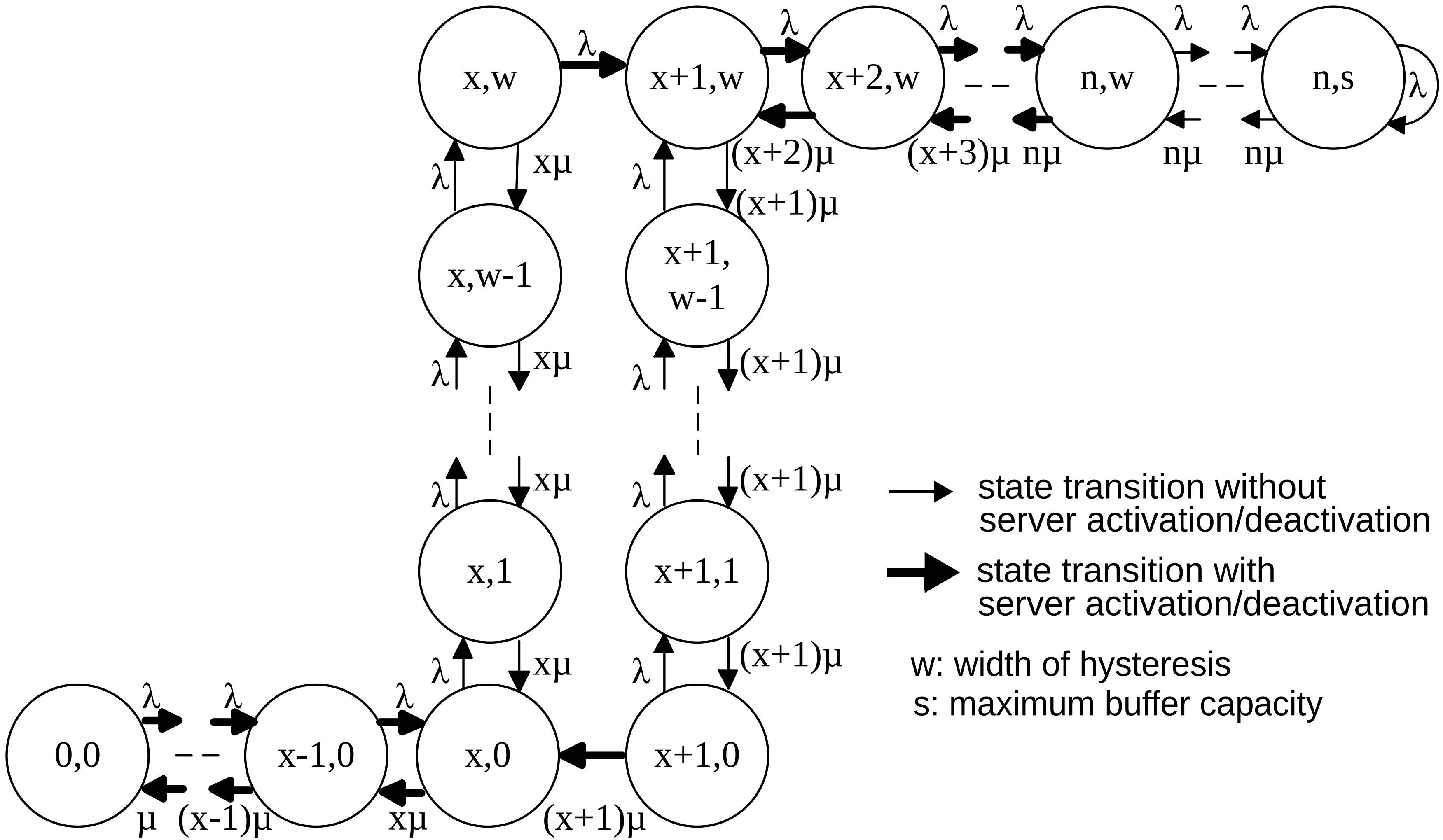
INDIVIDUAL DC MODEL



2.1-4 Modeling Algorithms

NON-ADAPTING MODEL BY FSM

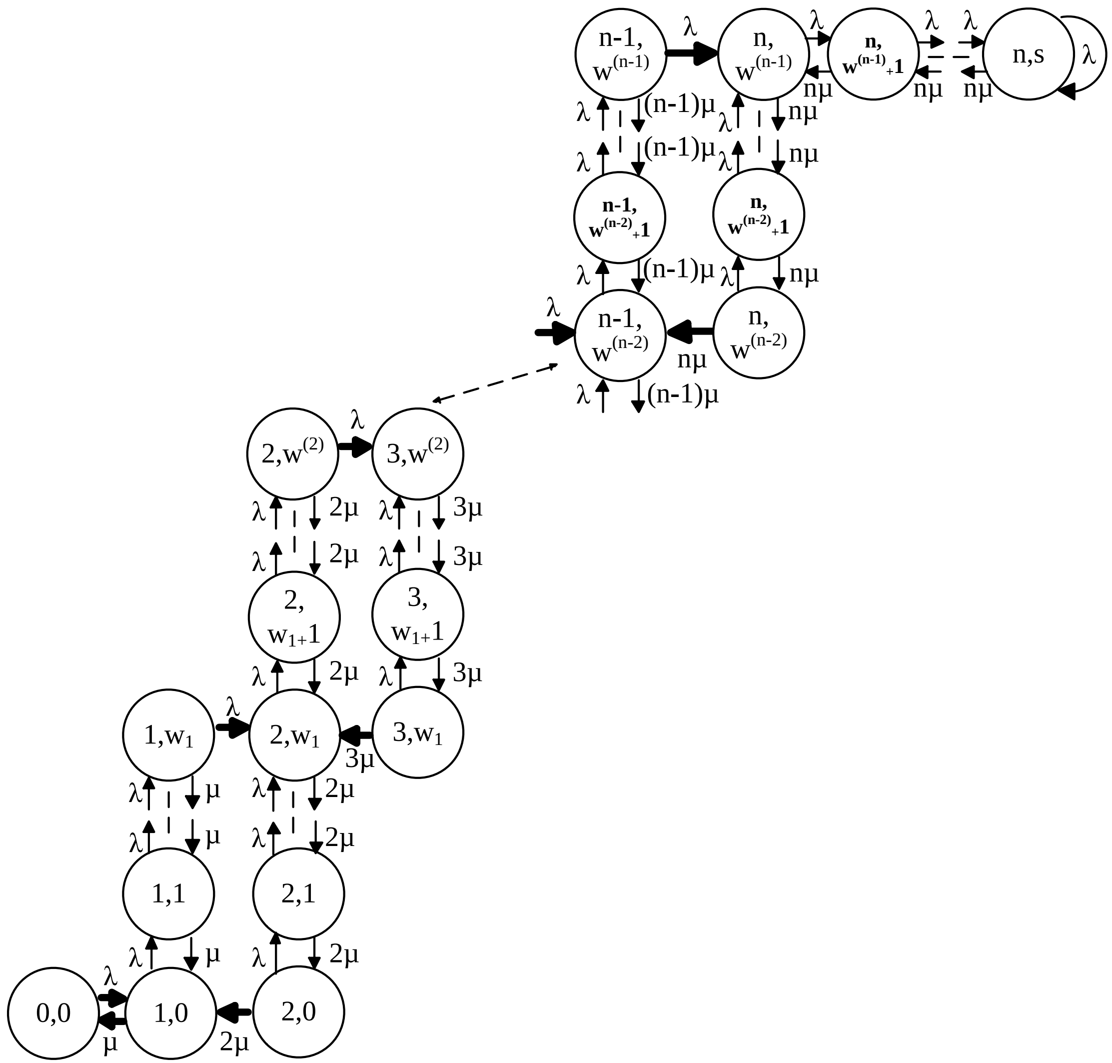
(1) SINGLE HYSTERESIS MODEL



2.1-5 Modeling Algorithms

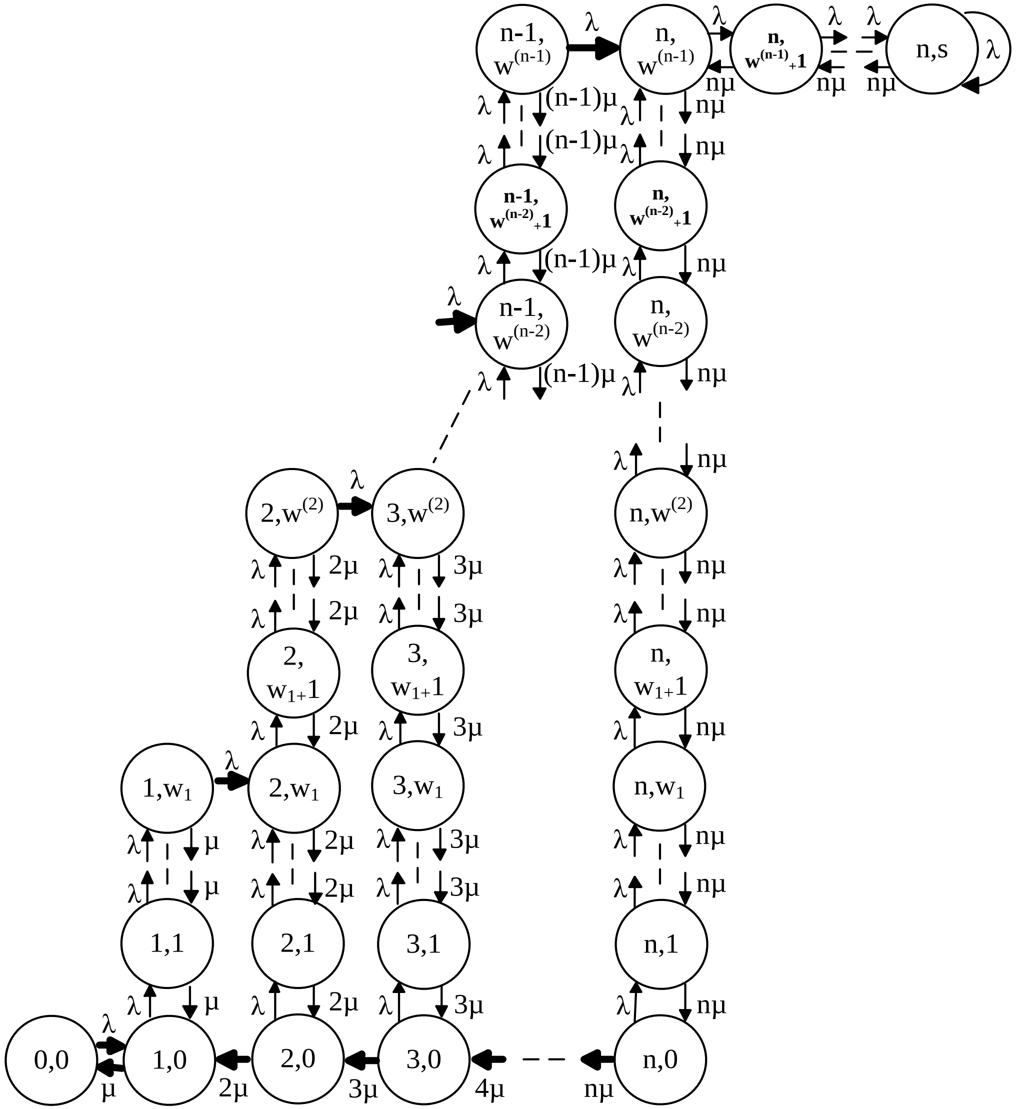
SELF-ADAPTING MODEL BY FSM

(2) MULTIPLE SERIAL HYSTERESIS MODEL



2.1-6 Modeling Algorithms

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



2.2-1 Performance Analysis and Results

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

2.2-2 Performance Analysis and Results

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

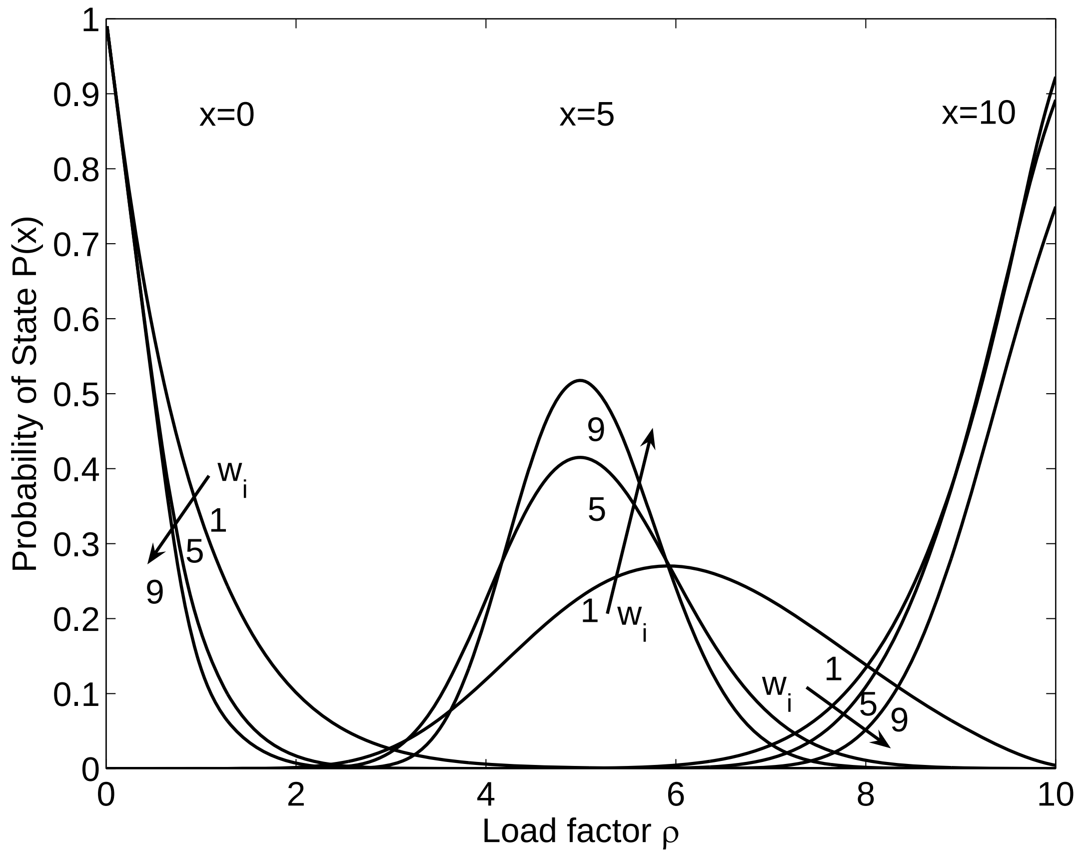
  In all solution methods certain generalizations are possible as

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

2.2-3 Performance Analysis and Results

NUMERICAL RESULTS (One DC only)

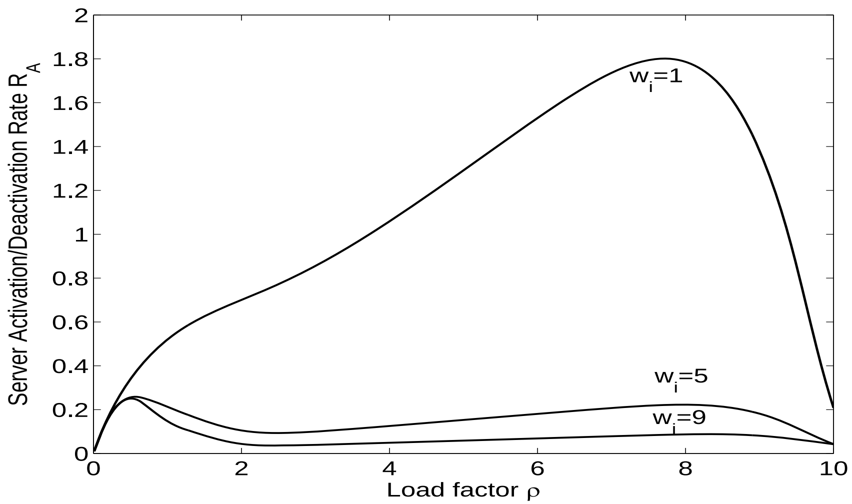
MULTIPLE SERIAL HYSTERESIS MODEL **Probabilities of State**



2.2-4 Performance Analysis and Results

NUMERICAL RESULTS (One DC only)

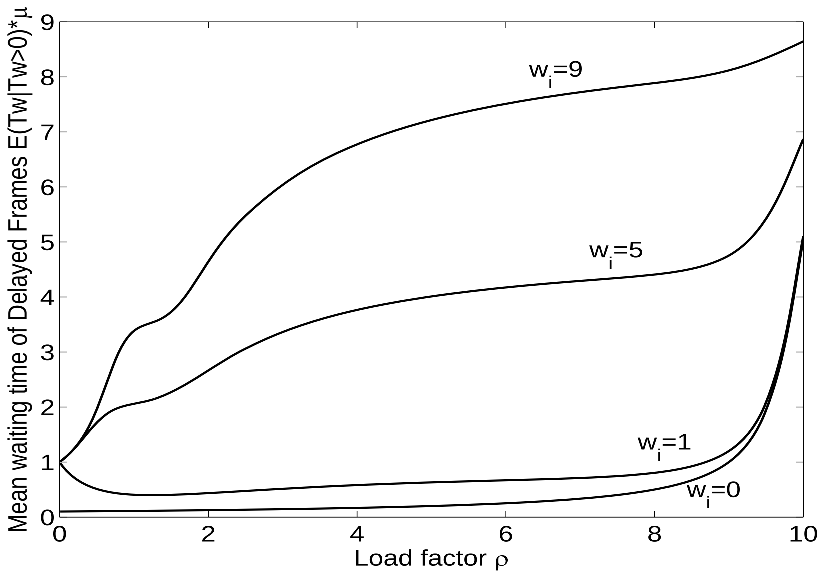
MULTIPLE SERIAL HYSTERESIS MODEL **Server Activation / Deactivation Rate**



2.2-5 Performance Analysis and Results

NUMERICAL RESULTS (One DC only)

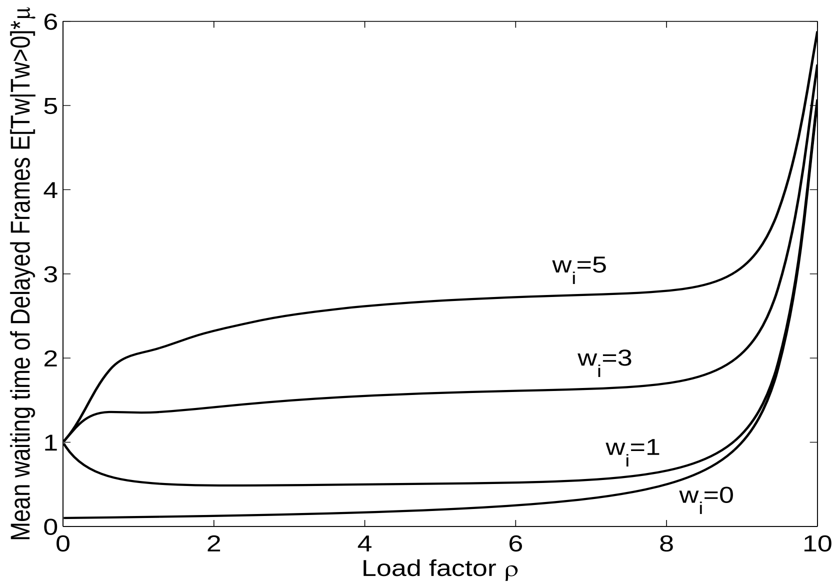
MULTIPLE SERIAL HYSTERESIS MODEL **Mean Waiting Time of Delayed Requests**



2.2-6 Performance Analysis and Results

NUMERICAL RESULTS (One DC only)

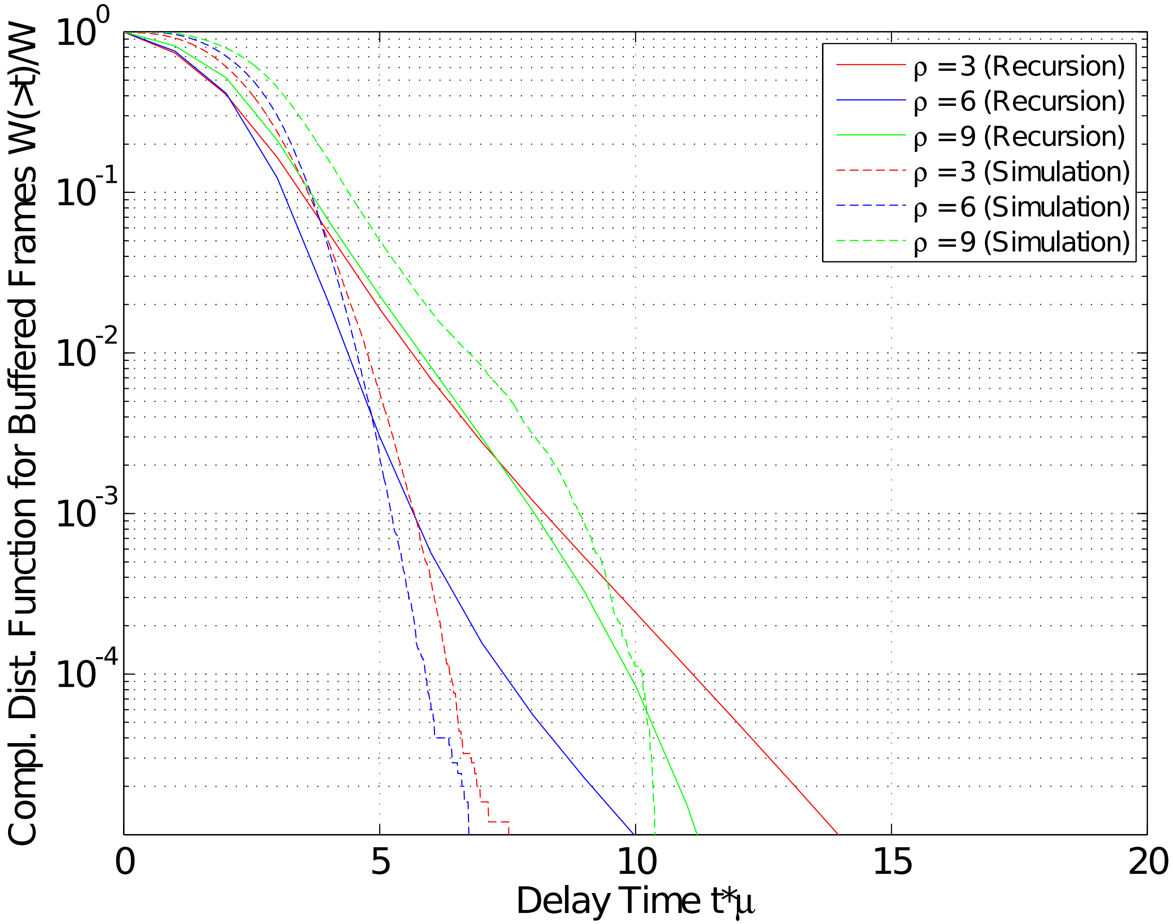
MULTIPLE PARALLEL HYSTERESIS MODEL **Mean Waiting Time of Delayed Requests**



2.2-7 Performance Analysis and Results

NUMERICAL RESULTS (One DC only)

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

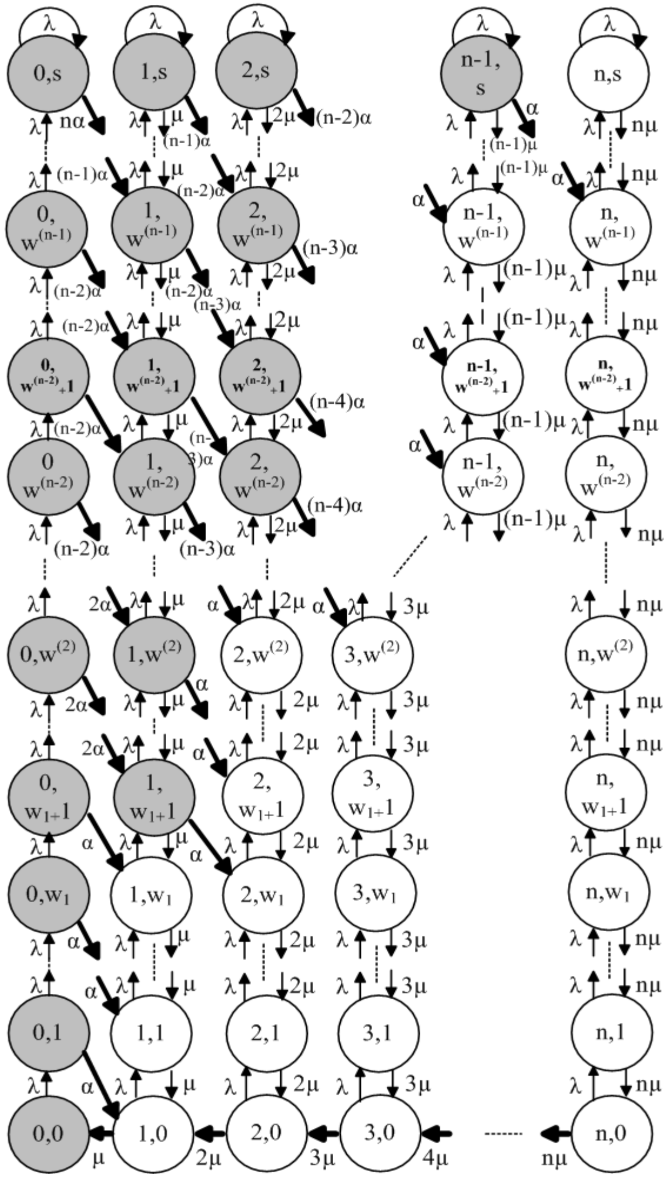


2.3-1 Modeling for Server Consolidation and Automatic Power Management

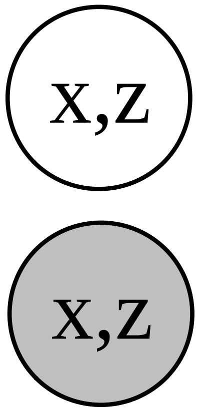
Conditions for the FSM Control:

- 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
- Guaranteed SLA either wrt average or percentiles of response time
- 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

2.3-2 Modeling for Server Consolidation and Automatic Power Management

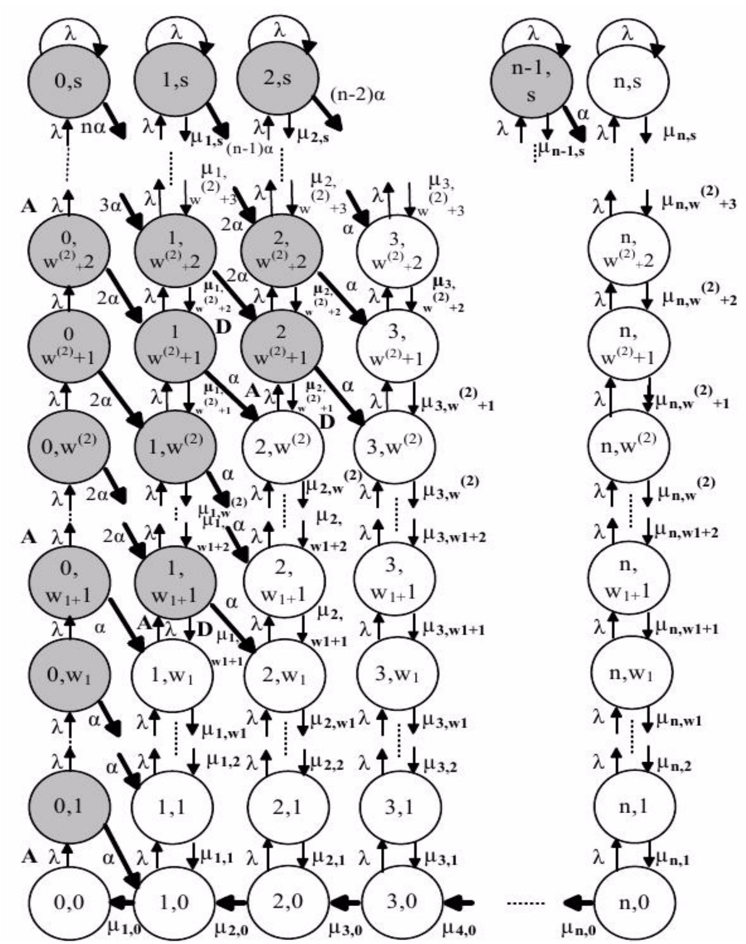


Multiple Parallel Hystereses
Multi-Server Queuing System
with/without Activation Overhead



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*

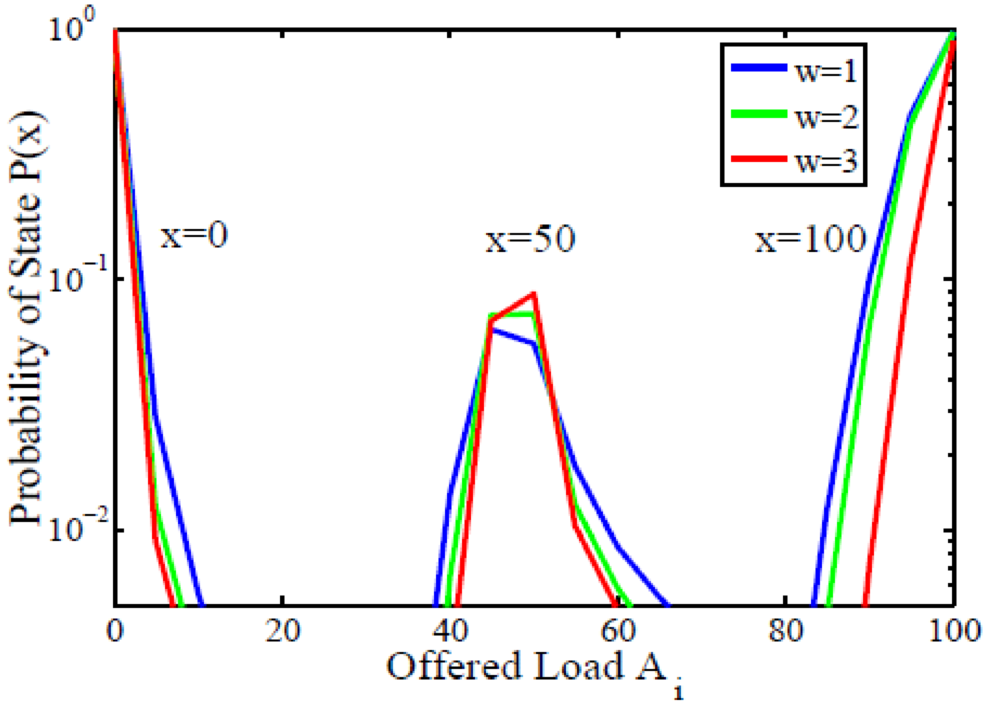


Figure: Probability of 'x' active servers vs. offered load
$n = 100$, $s = 300$, $\alpha = 1$, variable w

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

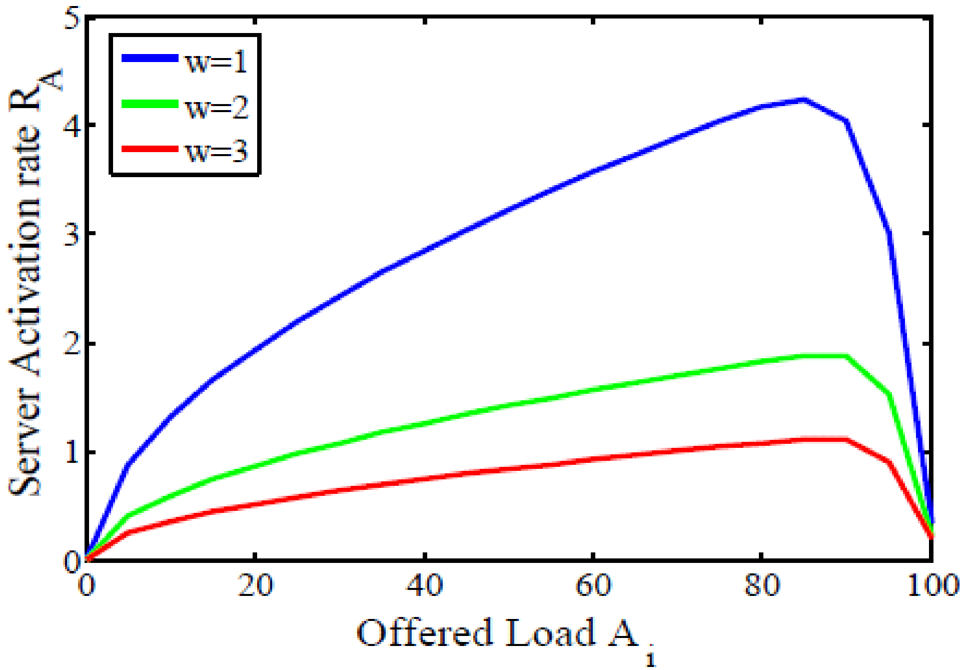


Figure: Server activation rate vs. offered load

n = 100, s = 300, α = 1, variable w

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

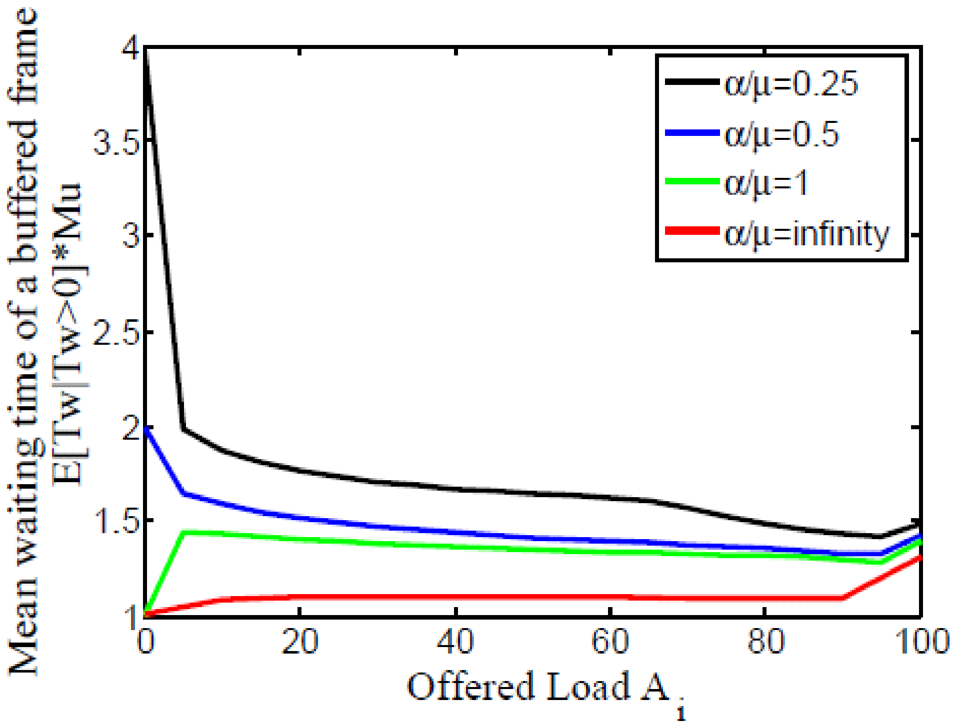


Figure: Mean delay of delayed frames vs. offered load

n = 100, s = 200, w = 2, variable α/μ

2.4-1 Load Balancing for Distributed Cloud Data Centers

STATIC LOAD BALANCING ALGORITHM

- Algorithm Steps
  1. Determine the maximum load that could be handled by each data center
     $A_{(max,i)} = [\text{function } (n_i) \mid t_w < t_{SLA}]$
  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):

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

2.5 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