Modeling and Analysis of Triple Play Services in Multimedia Internet

Natalia M. Markovich
ICS, Russian Academy of Sciences, Moscow

Tien Van Do
Budapest University of Technology and Economics, Budapest

Udo R. Krieger
Otto-Friedrich Universität, Bamberg

HET-NETs 2010
Zakopane, January 2010

Outline

- Part I: Service Architecture of Multimedia Internet
  - Infrastructure of a multimedia Internet
  - Protocol structures of multimedia services
  - Resource management and multi-processing strategies
- Part II: Modeling and Analysis of an Apache Web Server
  - A queueing network describing a dynamic pool of service processes
  - Workload modeling and performance analysis at the page layer
- Part III: Service Analysis of a Media Server
  - Performance modeling at the session level by a multi-class loss system
- Part IV: Analysis of Packet-Switched Multimedia Streams
  - Characterization of multimedia packet streams by measured data
  - Statistical modeling of Qos indicators by a bufferless fluid approach
  - Capacity requirements of peer-to-peer and VBR packet traffic
- Part V: Conclusions and Open Issues

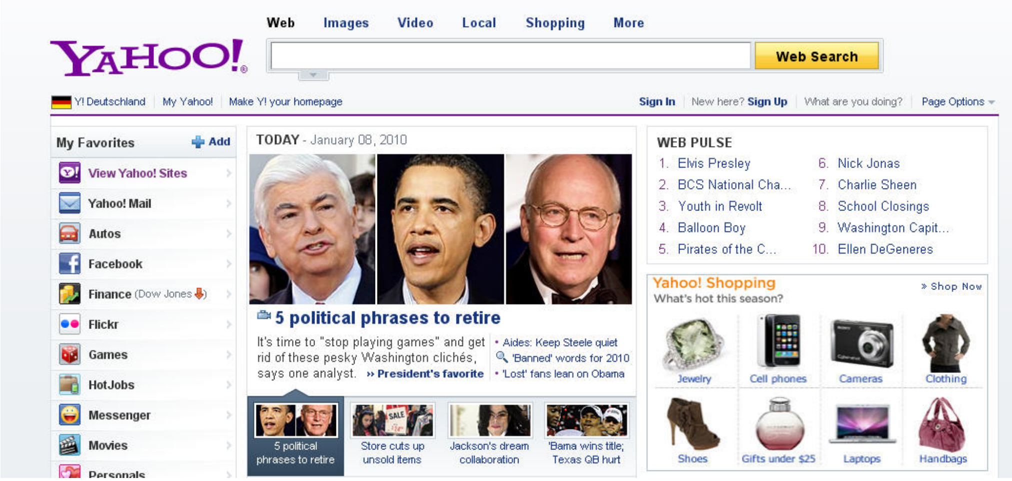
Part I

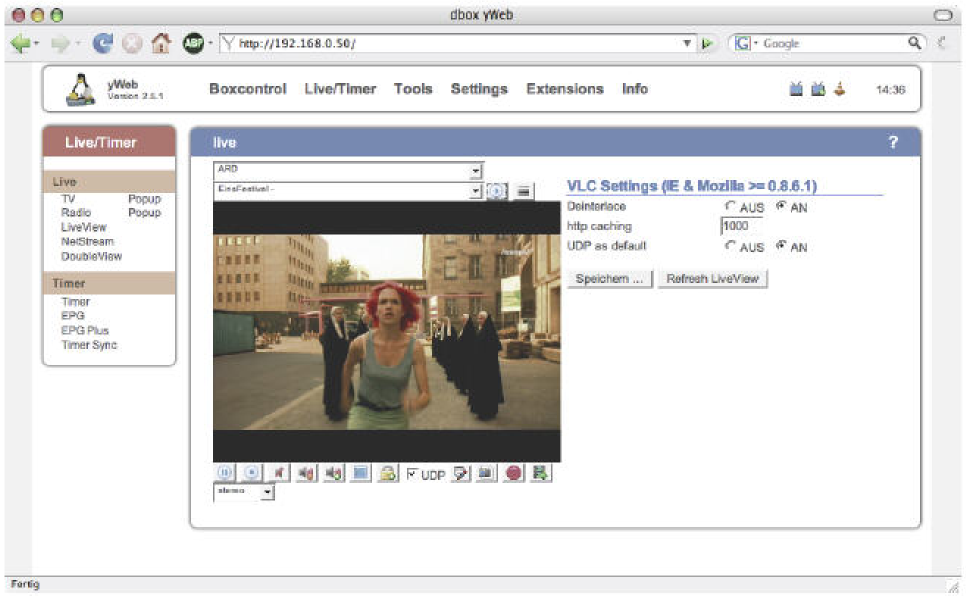
I. Service Architecture of Multimedia Internet

Part I

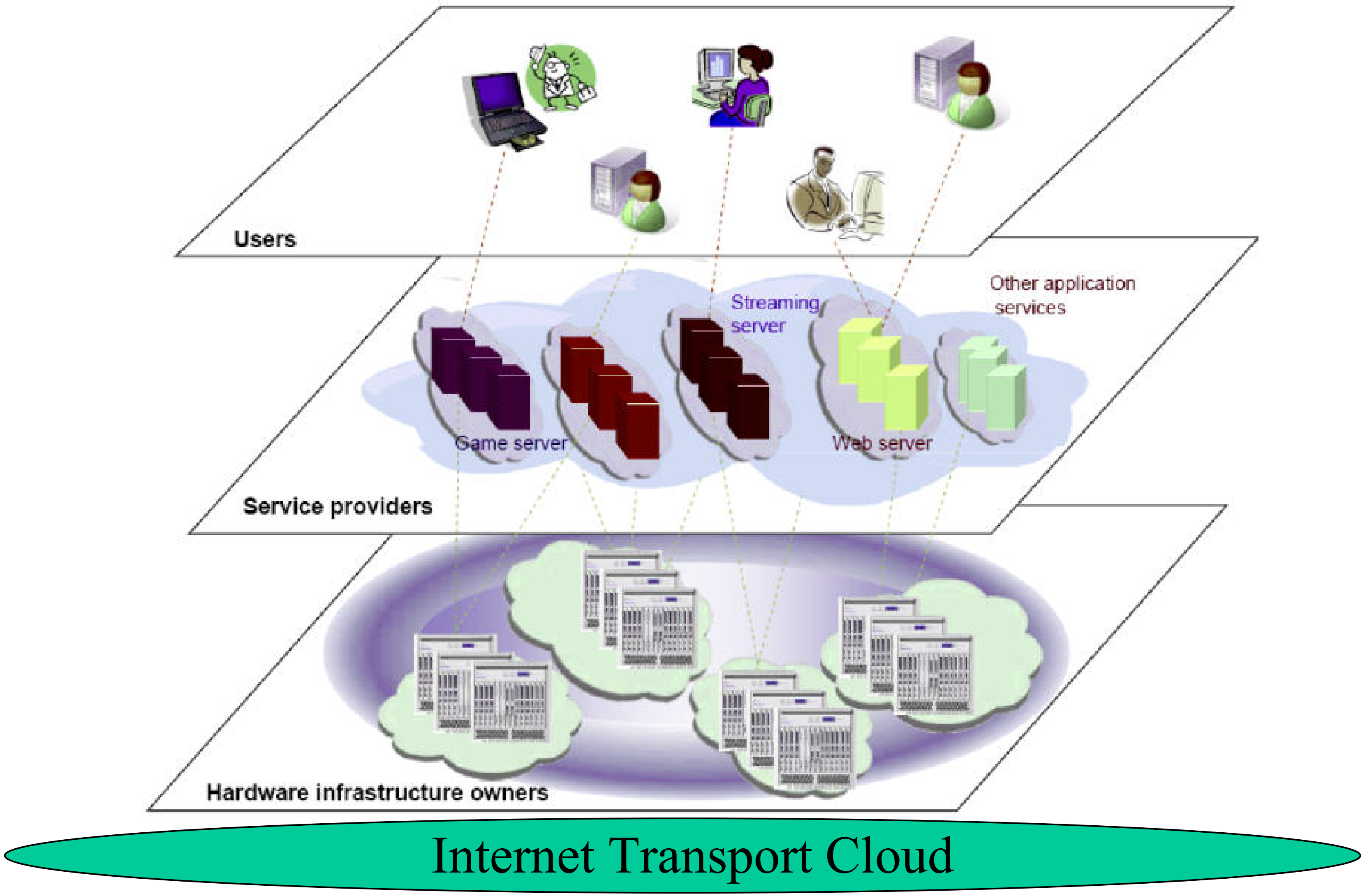
I.1 Infrastructure of a Multimedia Internet

Triple Play Applications

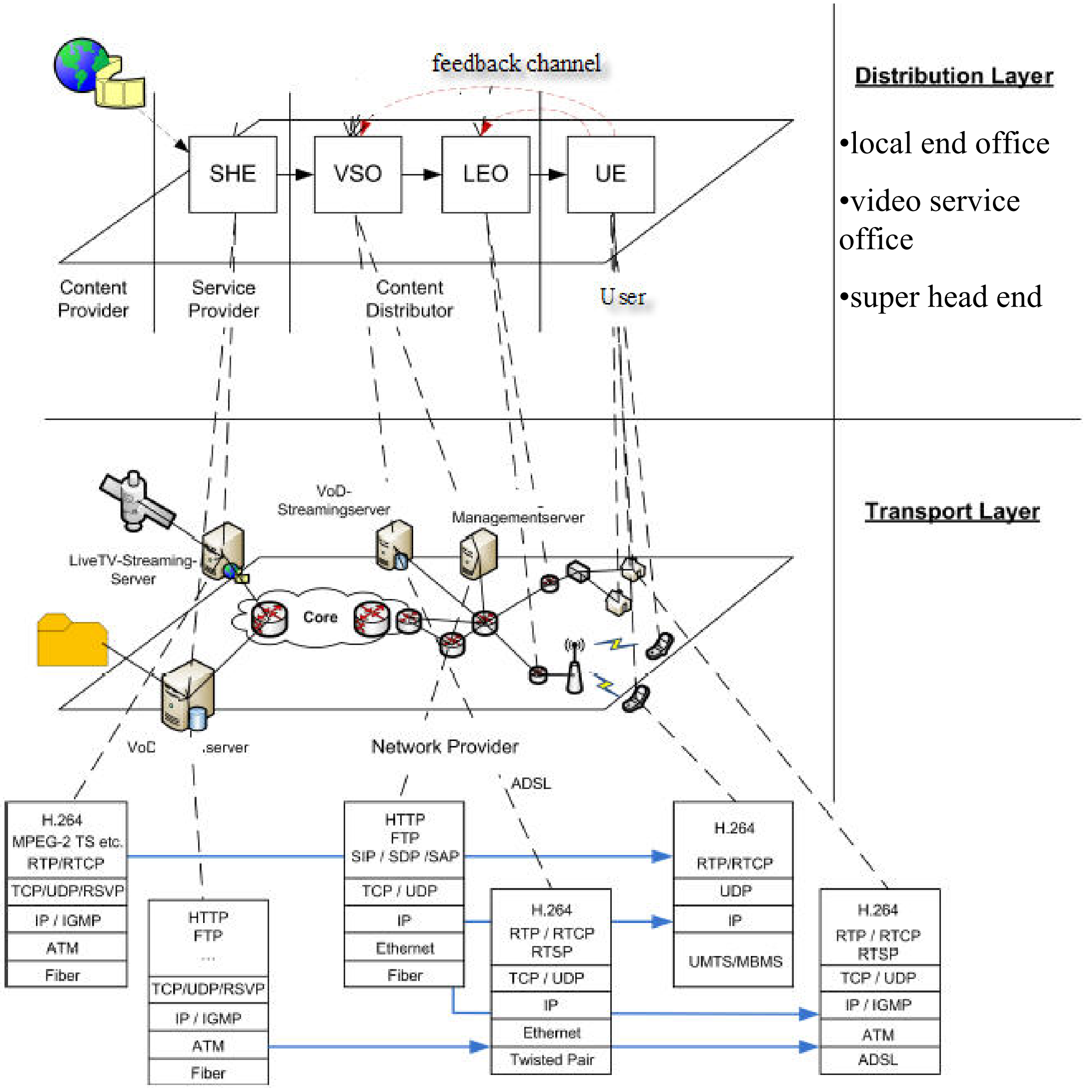




Hierarchical Structure of Service Architectures



Network Structure and Layering in IPTV Architectures



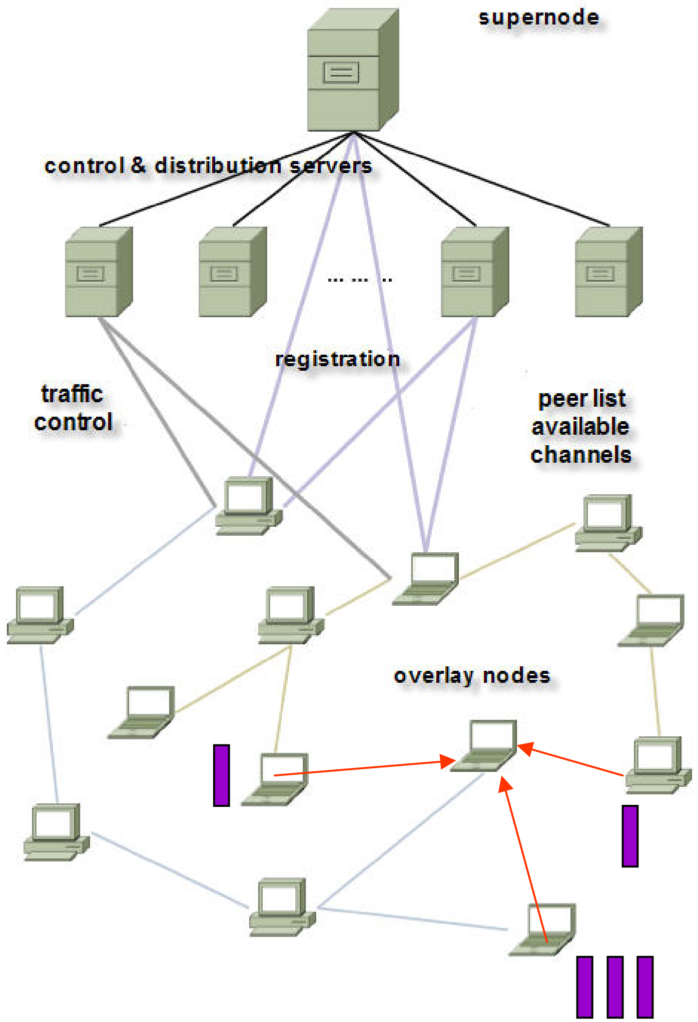
Distribution Layer

- dissemination of content
- provided by an overlay infra-structure

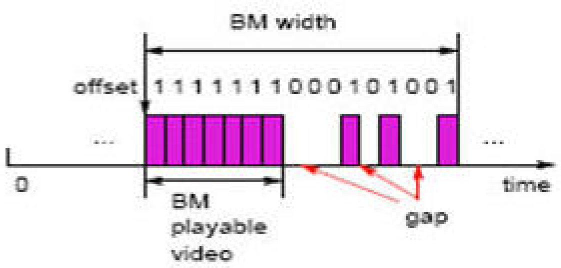
Transport Layer

- physical transport infrastructure with optical signal transport in the core and close to the homes
- on top of a packet-switched network
- use of RTP/RTSP combined with QoS supported TCP/UDP/IP stack
- SIP signaling applied

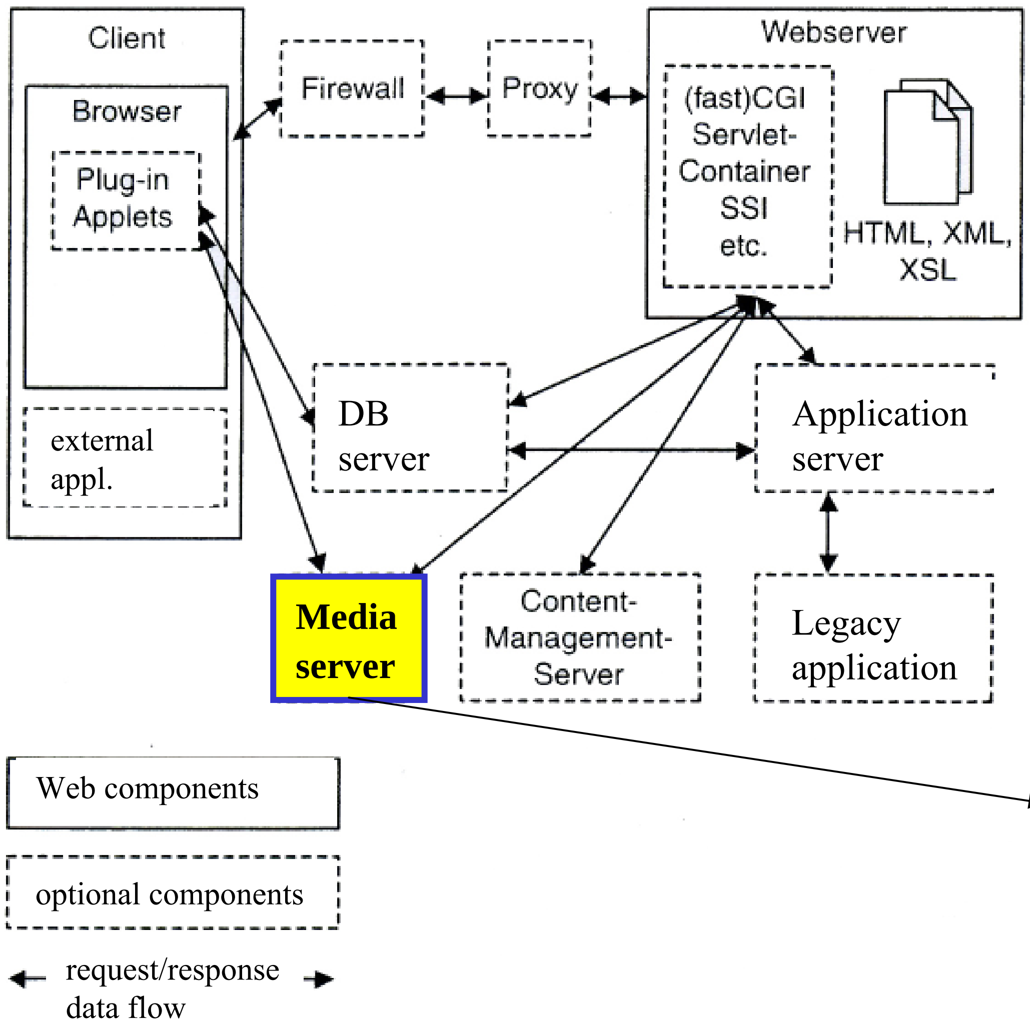
Peer-to-Peer Overlay Architecture



- Server-asssisted P2P Joost architecture
  - mesh overlay topology
  - m:n unicast connections
- video encoding by MPEG-4/H.264 AVC
  - very good quality of video transport
  - due to error resilience of the codec
  - good on all resolutions of 720 x 576 up to 1280 x 1024
- Applied Protocols
  - HTTPs (TSLv1) for registration and authentication
  - UDP for the transport of chunked video data indicated by bitmaps



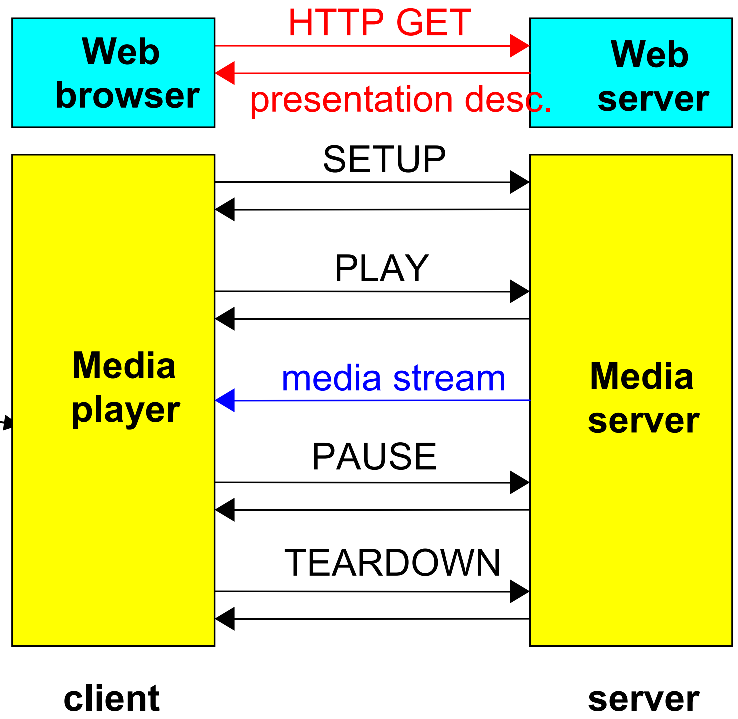
Simplest Multi-Tier Web Service Architecture



Client-Server Model

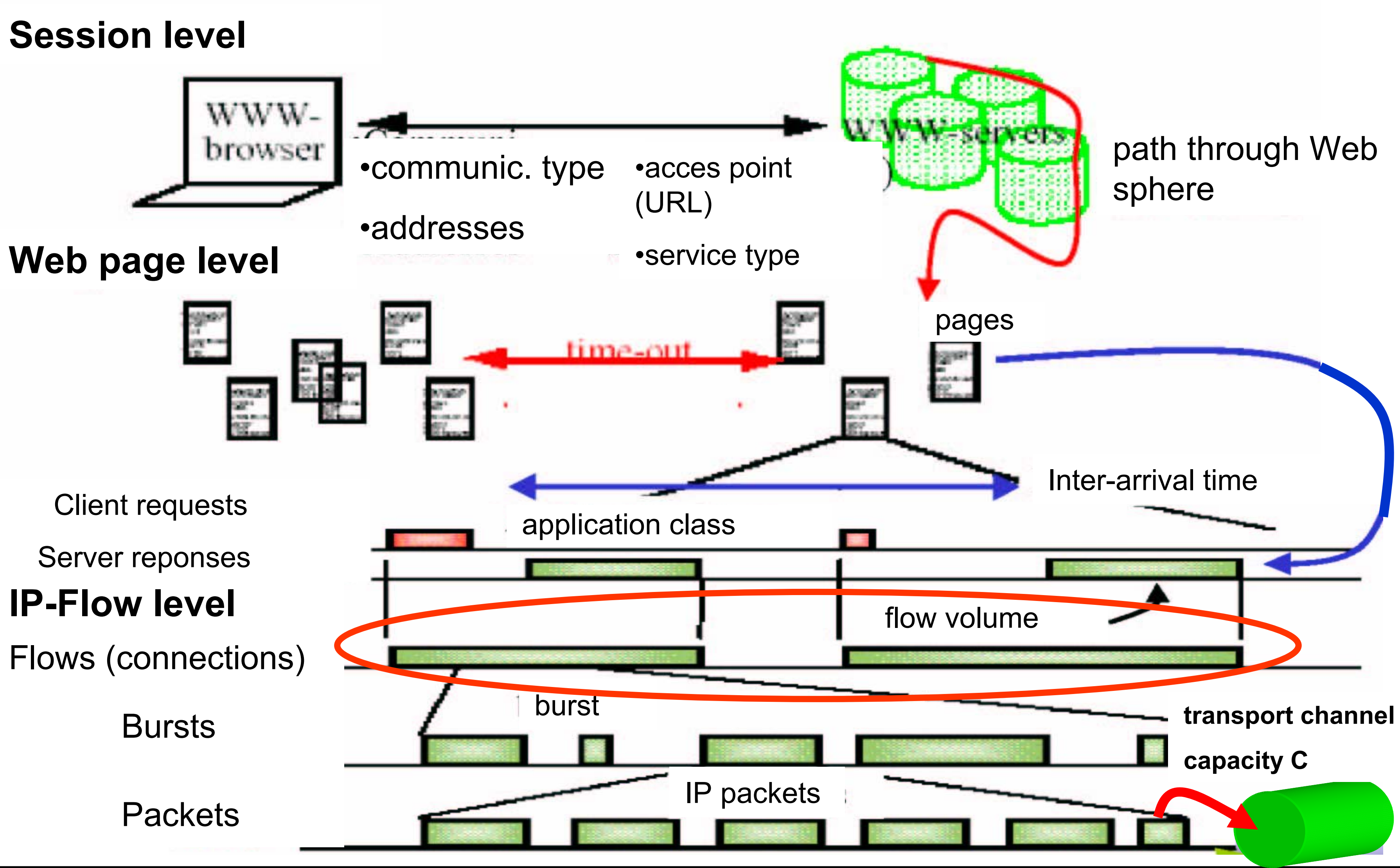
- as entrance gateway to Web services

- multimedia service architecture

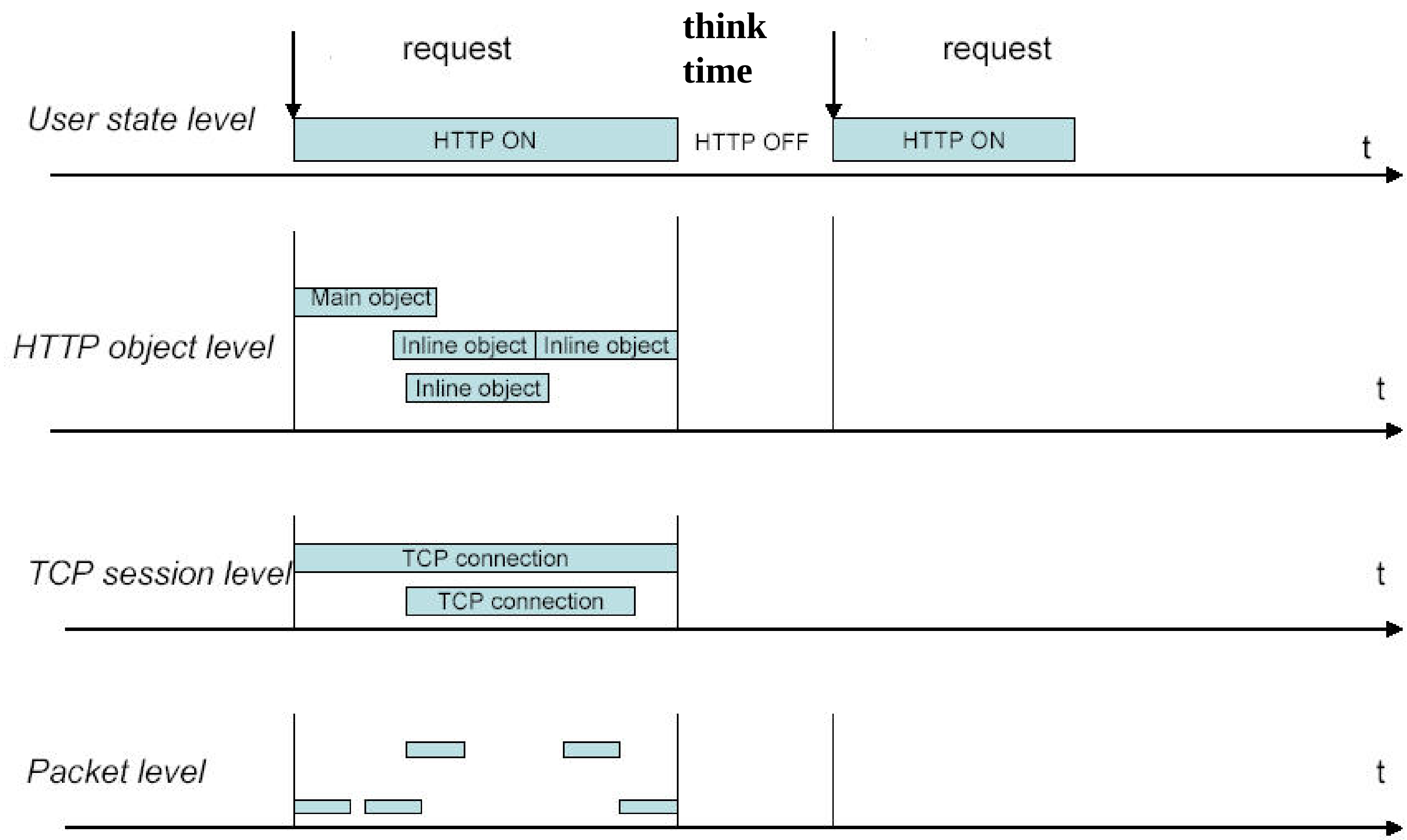


I.2 Protocol Structures of Multimedia Services

Structure of Client-Server Traffic in Internet

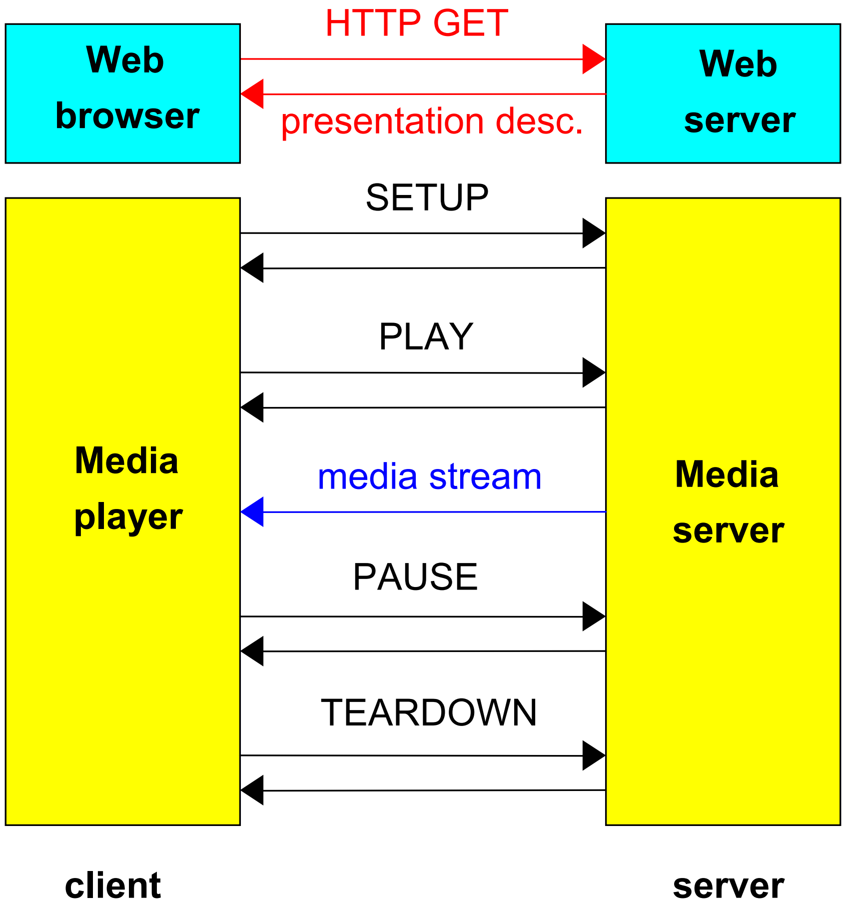


Hierarchical Workload Modeling of Web Traffic



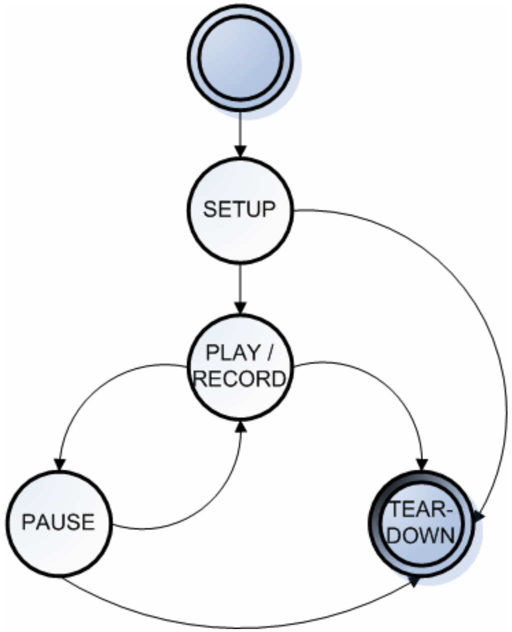
Real-Time Protocol Stack: RTP and RTSP

Multimedia Service Architecture



RTSP

- initializes, controls and stops the streaming
- transmitted over TCP
- description of content via SDP
- operation controlled by a state model
- allocates ressources



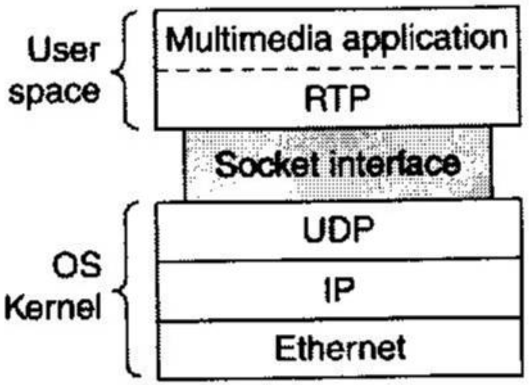
Real-Time Protocol Stack: RTP and RTSP

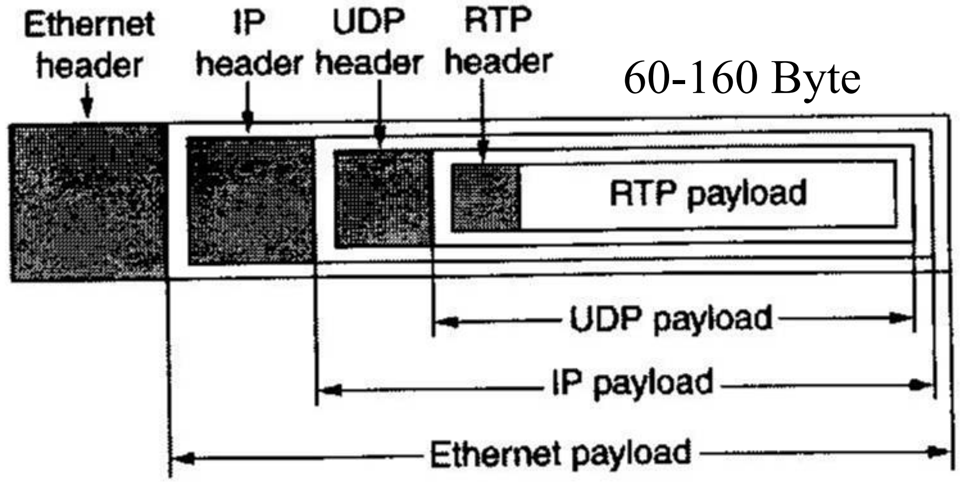
- RTP runtime library provides a transport layer interface
  - is an extension of UDP
- API provides :
  - UDP port numbers, IP addresses
  - error control on segments
  - identification of payload data types
  - sequence numbers of packets
  - time stamps

Overhead: 54 Byte

14 20 8 12 Byte

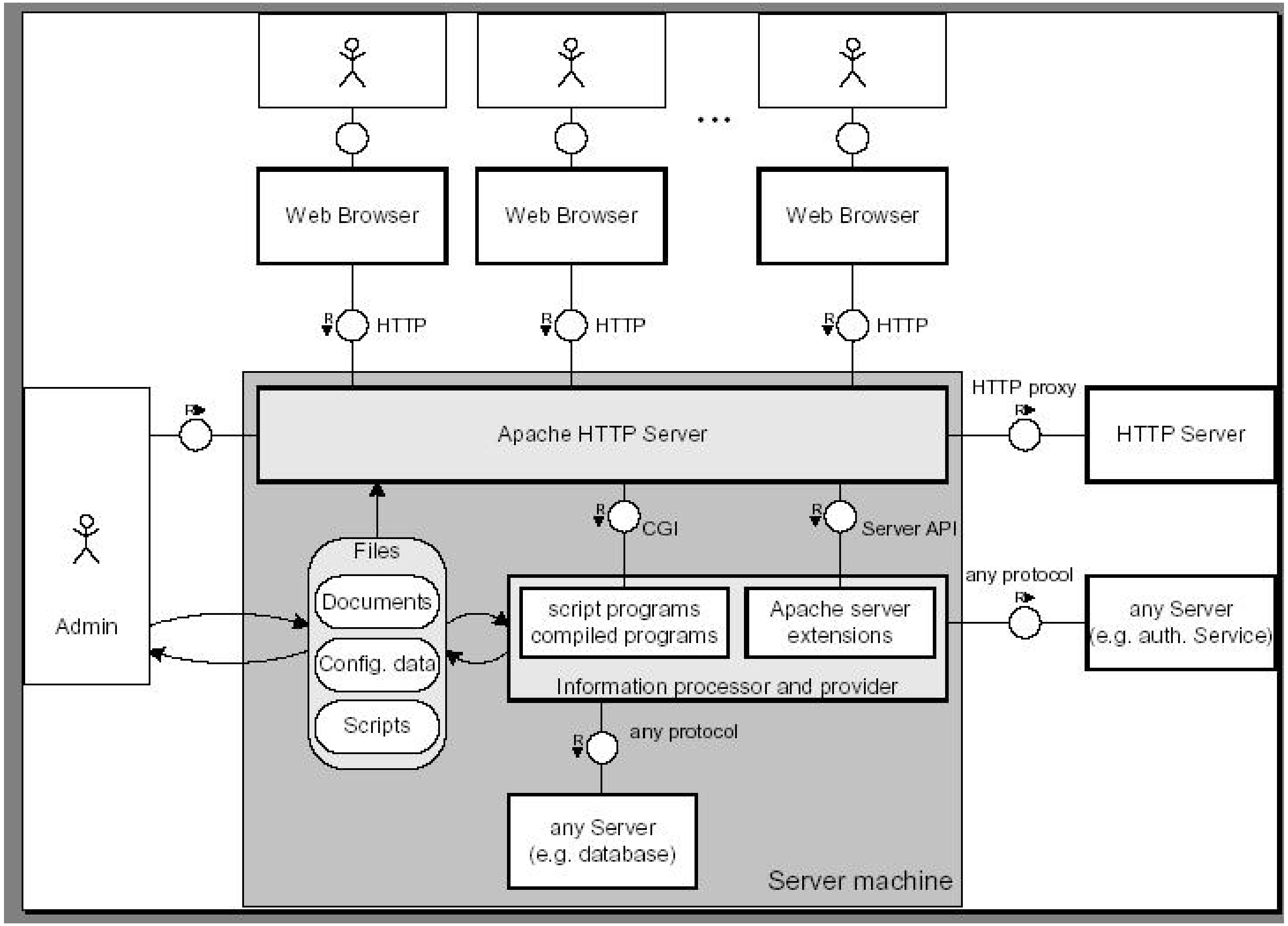
60-160 Byte





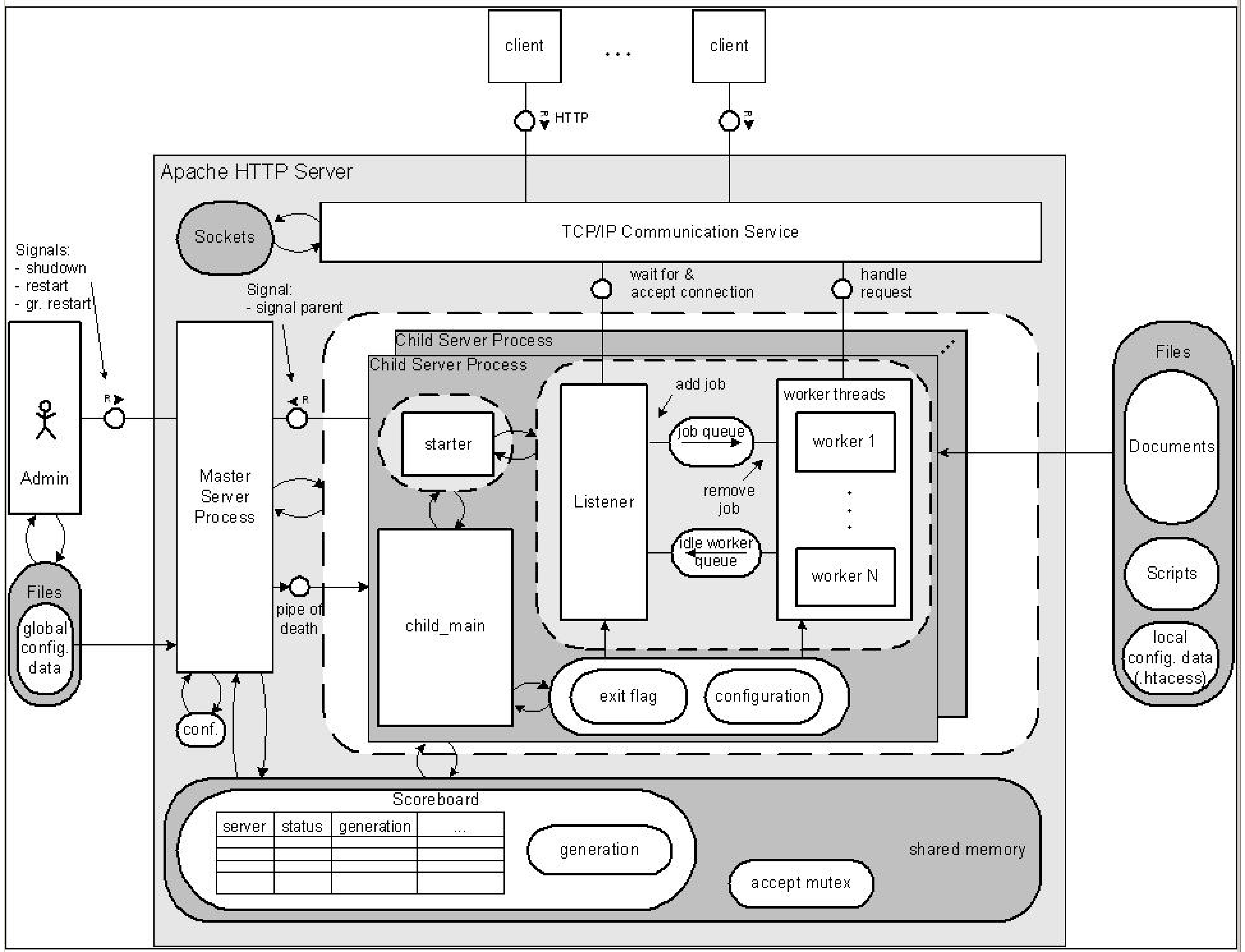
I.3 Resource Management and Multi-Processing Strategies

Structure of an Apache Web Server



source: The Apache Modeling Project, http://apache.hpi.uni-potsdam.de

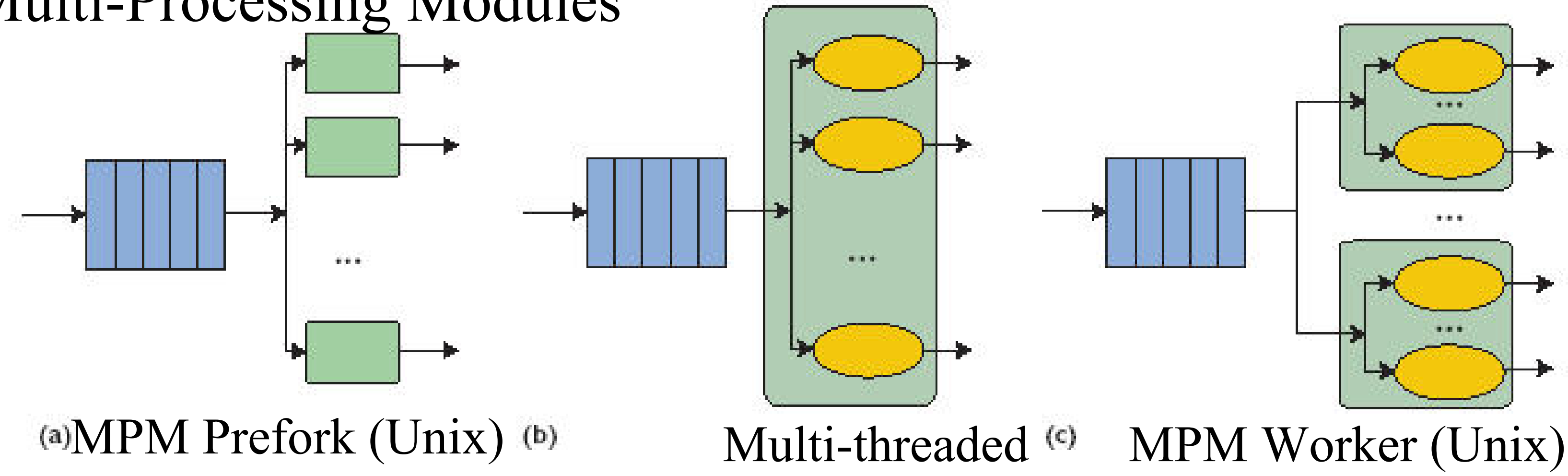
Multi-Processing Modules and Worker Threads Processing

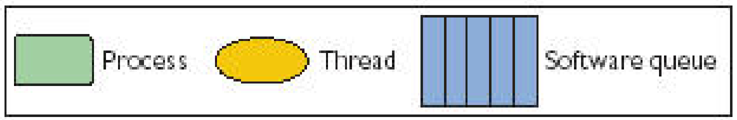


source: The Apache Modeling Project, http://apache.hpi.uni-potsdam.de

Multi-Processing Operation of an a Web Server

Multi-Processing Modules





•how many threads/processes serve requests?

Resource management & control by Apache directives:

•*MaxClients* (N): - limits the maximal number of simultaneous connection requests

•*MaxSpareServers* (h_{max}): - limits the maximal number of idle HTTP service processes,
- exceedance $\Rightarrow$ killing of idle processes with rate ε

•*MinSpareServers* (h_{min}): - limits the minimal number of idle service processes,
– dropping $\Rightarrow$ creation of new processes with rate η

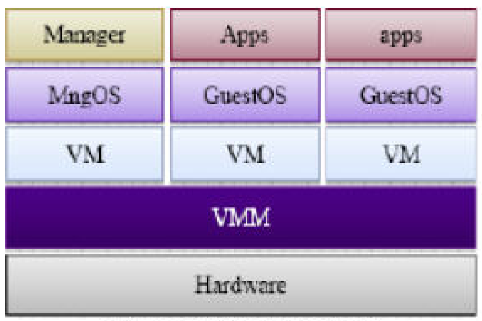
Virtualization of Server Processing

• virtualization of task processing, resource assignment and management in current operating systems

• a virtualization layer (VMM/VM) separates physical hardware and resource demands of application processing

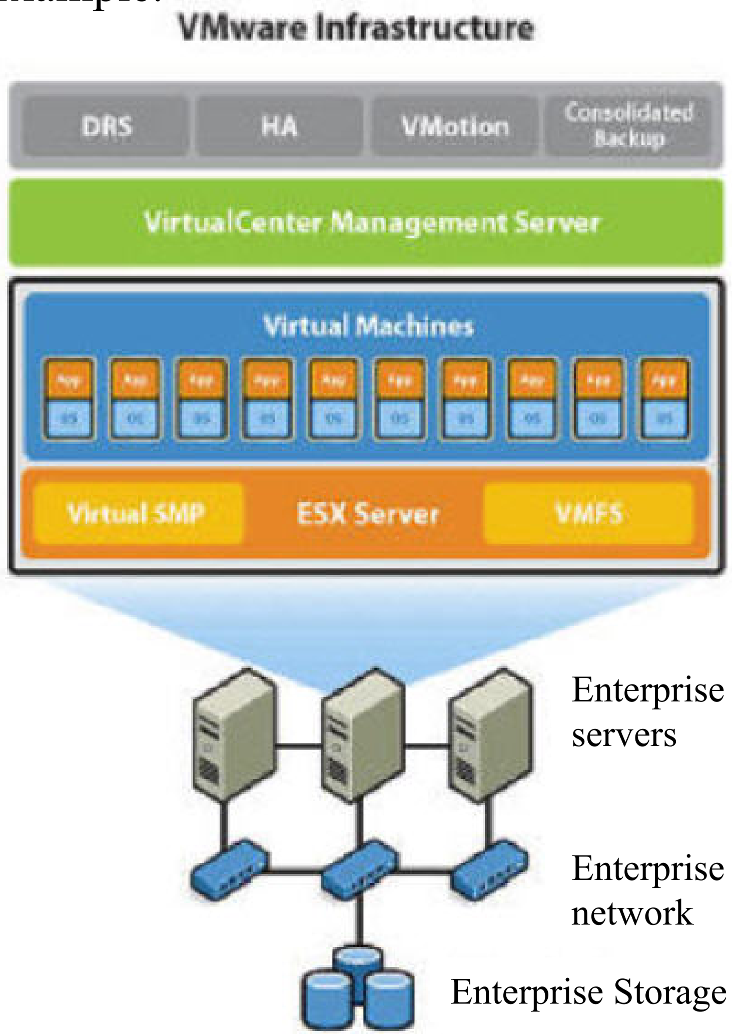
• integration of different operating systems (guest OS)

  • running on one physical system (host OS) ⇒ server consolidation



Hypervisor-based virtualization

Example:

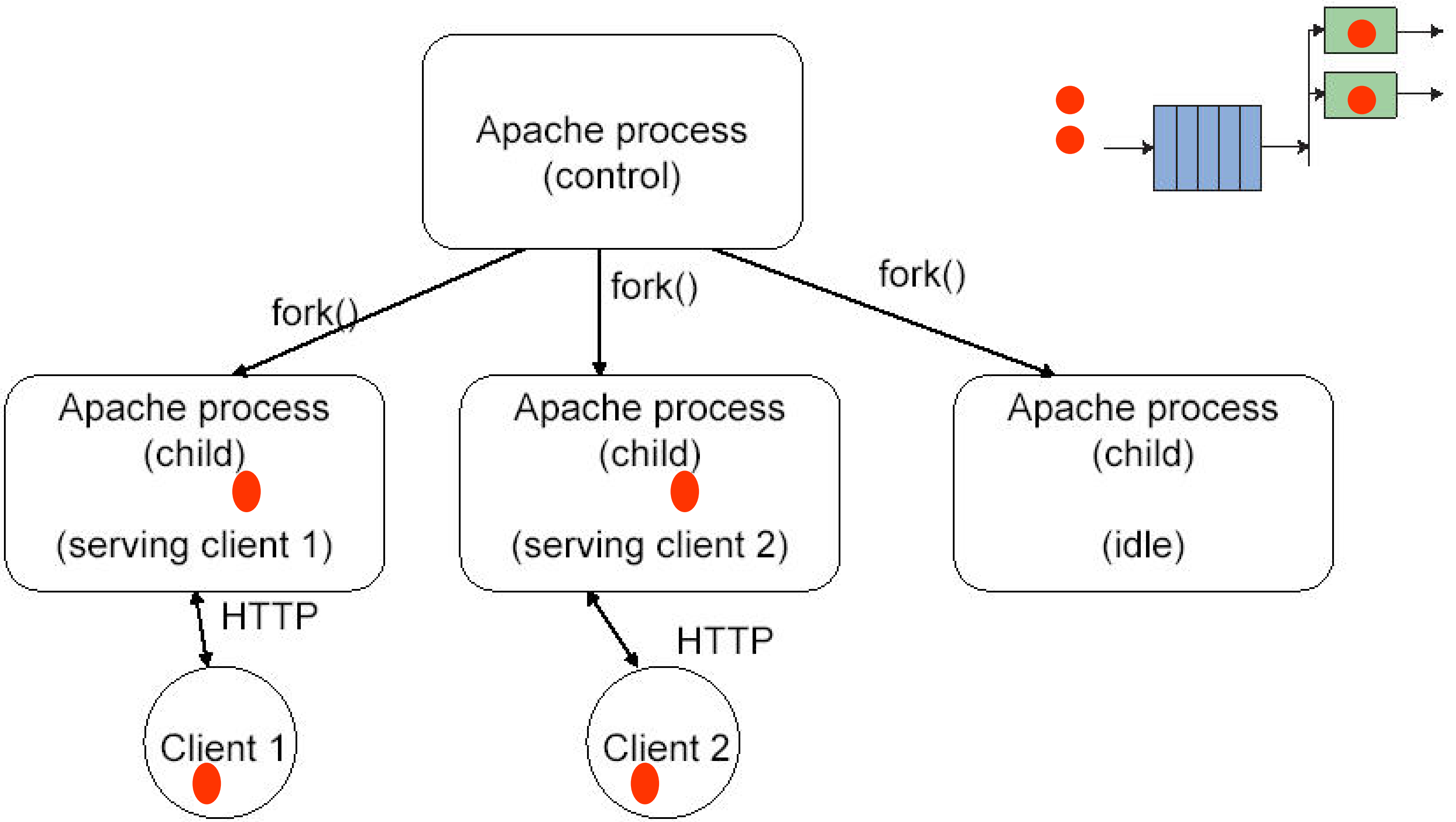


Part II

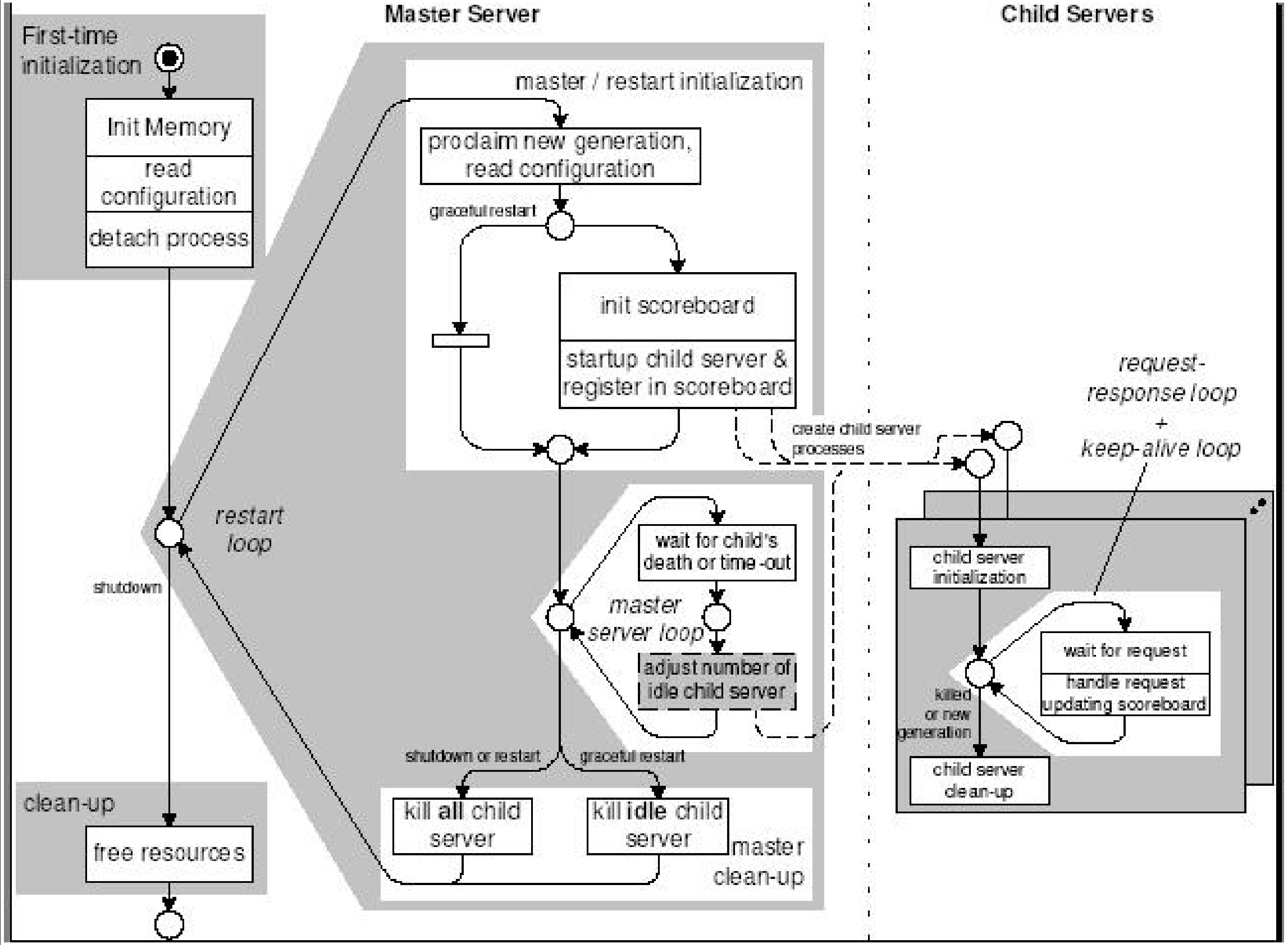
II. Modeling and Analysis of an Apache Web Server

II.1 A Queueing Network Describing A Dynamic Pool of Service Processes

Apache's Multi-Processing Module Prefork

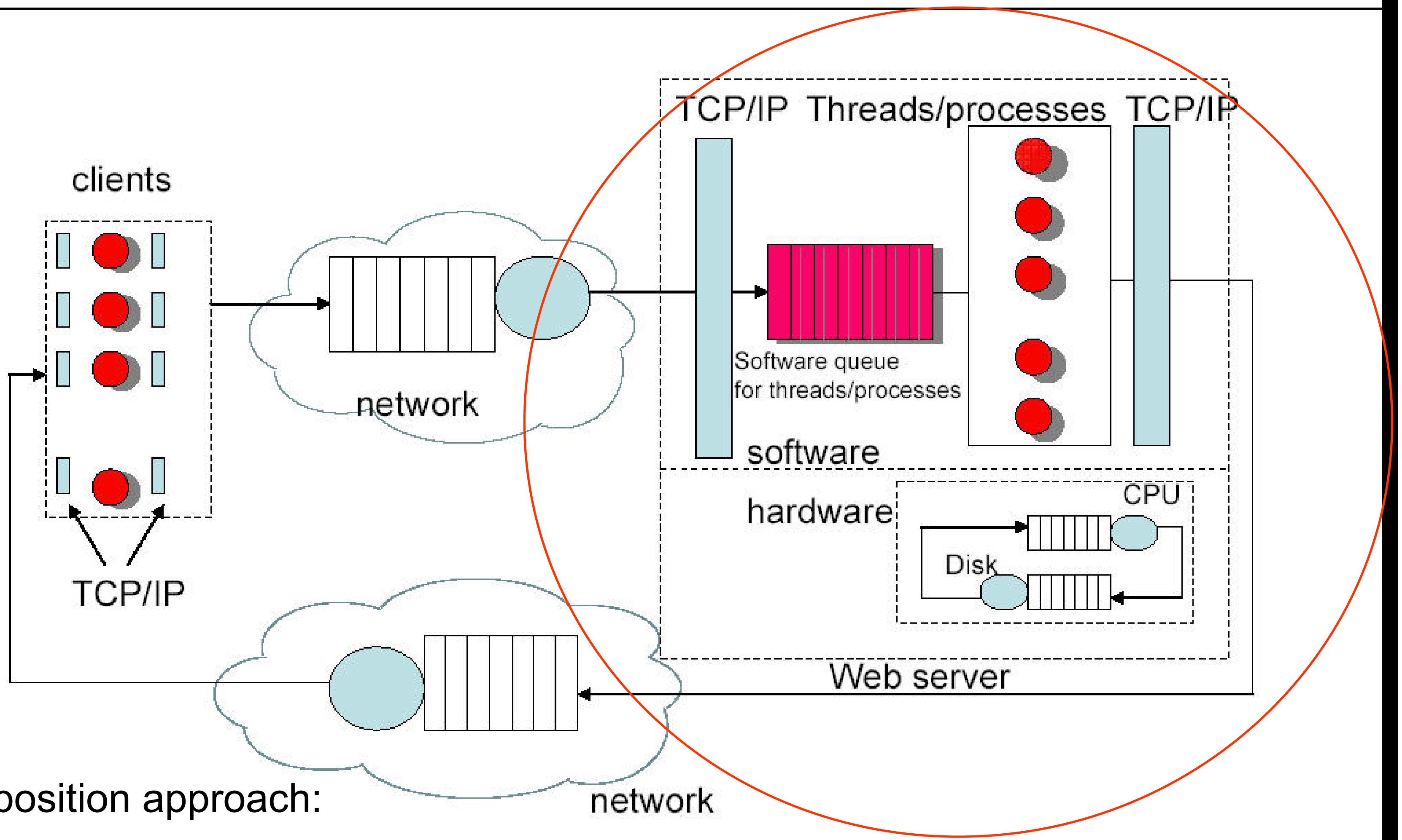


Concept of Master-Slave Processing by Apache



source: The Apache Modeling Project, http://apache.hpi.uni-potsdam.de

Queueing Model of an Apache Web Server



Decomposition approach:

- isolate the Web server and model its workload
- server separable from the transport network and impairments of the TCP stack

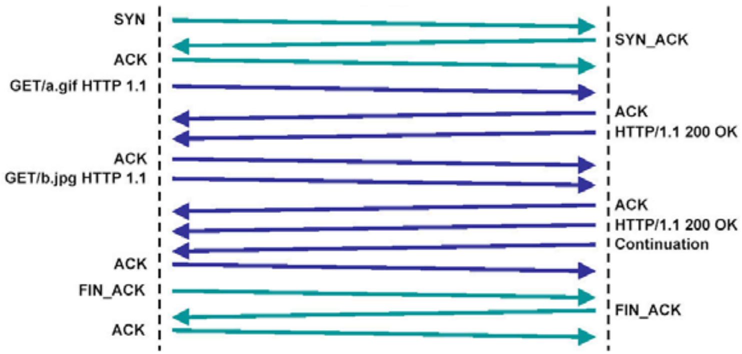
Part II

II.2 Workload Modeling and Analysis at the Page Layer

Liu´s Workload Modeling at the Page Layer

WAGON workload

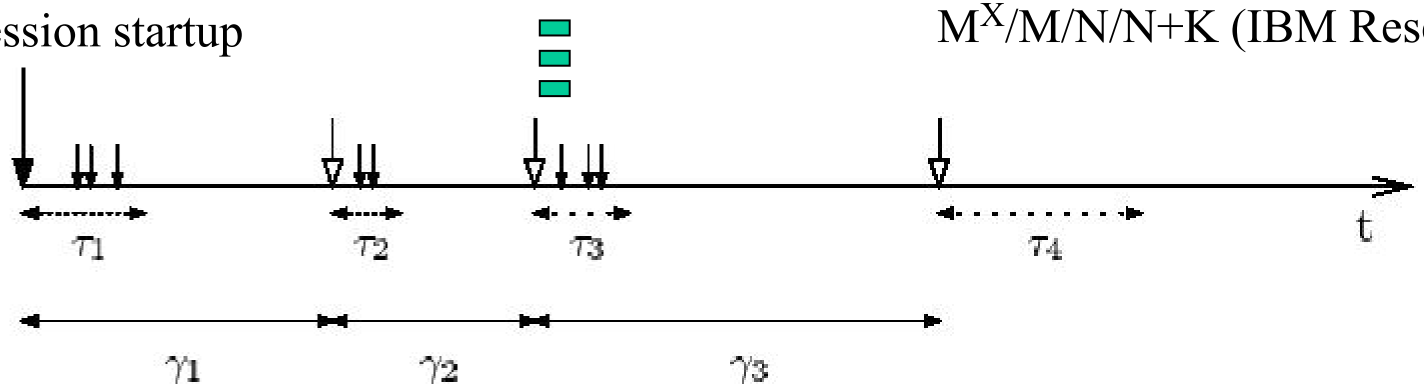
- download time of page τ_j
- time elapsed betwen each click γ_j
- request of an embedded object
- click j



Potential queueing model:

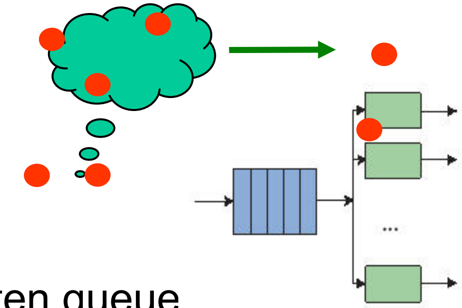
$M^X/M/N/N+K$ (IBM Research)

Session startup



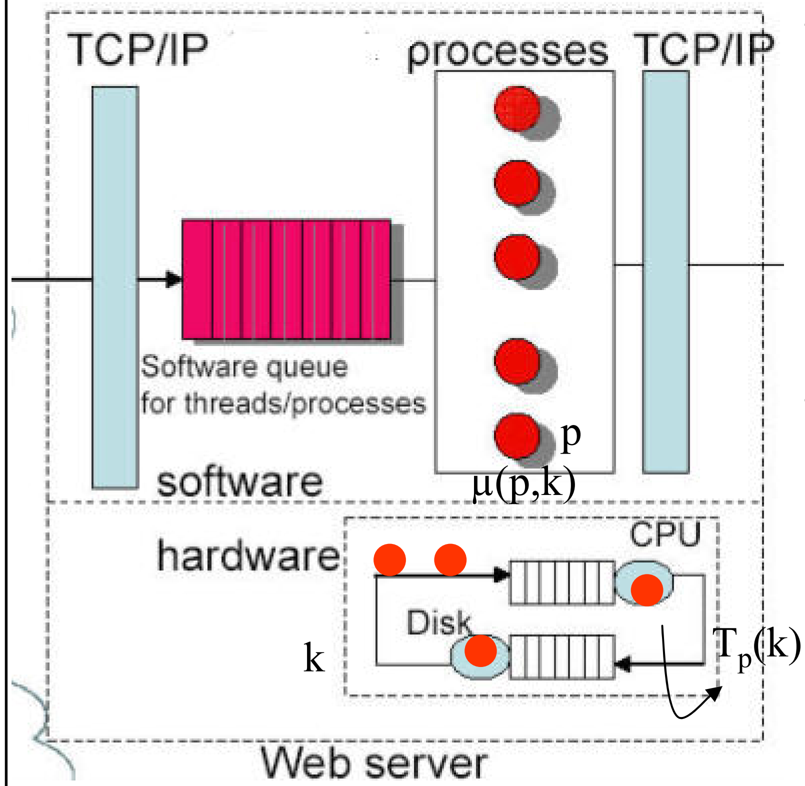
Our extension: $MAP^X/M/N/N+K$ system with *correlated Markovian* batch arrival process

Level-Dependent Matrix-Geometric Model



- System model $Z(t) = (I(t), J(t))$
  - CTMC on finite state space $\{1,.., M\times N\} \times \{0,..,L\}$
- *level J(t)*
  - number j of TCP sessions served or waiting in the listen queue
- *phase* $I(t) = F(I_1(t), I_2(t))$ determined by
  - modulator variable $I_1(t) = i \in \{1, \ldots, M\}$ of a *Markov-modulated batch arrival* stream with M states (compound Poisson process)
    - having a generalized exponential (GE) interarrival time τ with distribution in phase i: $\Pr(0 < \tau_i < t) = (1-\theta_i)(1-e^{-\sigma_i t})$
    - random batches S of size s with distribution $(1-\theta_i)\theta_i^{s-1}$
  - random number $1 \le I_2(t) = h \le N$ of available HTTP service processes
    - governed by *exponentially distributed* creation and killing times of pooled servers with rates η and ε

State-Dependent Service Time Modeling



- service time as result of complex interaction
  - of the contention at physical resources: CPUs, disk operations
  - TCP flow control and segment transfer in the Internet
- application of Menasce´s QNW model
  - aproximation of the *state-dependent service rate $\mu(p,k)$* by the throughput $T_p(k)$ of a closed QNW
  - given k requests served by p active processes in the system
  - $T_p(k) = T_p(p), \quad k \geq p$
- applied alternative:
  - *identification by measurements*
- **compromise: accuracy vs. complexity**

Part II

II.3 Performance Analysis of an Apache Server

Performance Measures

- mean number of idle service processes

$$E(Idle) = \sum_{i=1}^{NM} \sum_{j=0}^{L} p_{i,j} \cdot \max(f_2(i) - j, 0)$$

- probability p_W that a new request is waiting for an idle child process

$$p_W = \sum_{i=1}^{NM} \sum_{j=0}^{L} \sum_{s=\max(f_2(i)-j,0)+1}^{\infty} p_{i,j} \cdot (1 - \theta_{f_1(i)})\theta_{f_1(i)}^{s-1} \cdot \frac{\min((s - \max(f_2(i) - j, 0), L - j)\sigma_{f_1(i)}}{\overline{\sigma}}$$

  - with mean arrival rate $\overline{\sigma} = \sum_{l=1}^{M} \frac{\sigma_l}{(1 - \theta_l)} r_l.$

  - and steady state vector r of the modulator I_1:

$$\mathbf{r} \cdot Q_X = 0 \quad ; \quad \mathbf{r} \cdot e_M = 1.$$

Validation of the Approach by Measurements

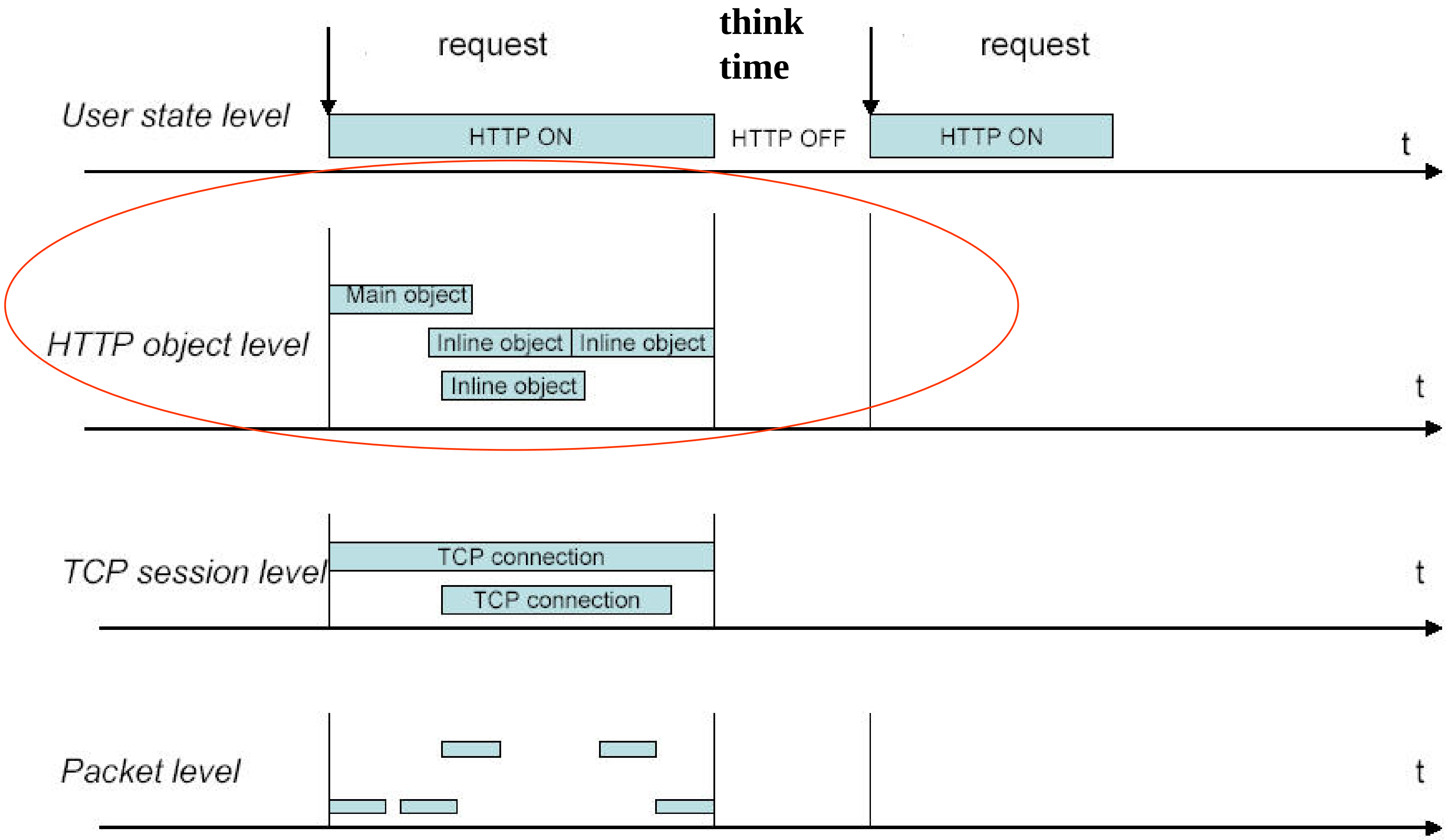
Parameters				Measurement Study			Analytic Model		
σ	θ	η	ϵ	E(Idle)	Var(Idle)	Waiting Prob.	E(Idle)	Var(Idle)	Waiting Prob.
5	0.4	0.003960	0.001500	9.92740	0.4323	0.000087	9.933774	0.333595	0.0001343
5	0.5	0.007042	0.003130	9.89174	0.6153	0.003466	9.906068	0.437426	0.0012225
5	0.6	0.011198	0.007337	9.83198	0.8833	0.015859	9.858125	0.602280	0.0075010
5	0.7	0.020465	0.016675	9.70886	1.3312	0.056009	9.762665	0.895090	0.0352133
5	0.8	0.062304	0.058661	9.47287	2.0765	0.175968	9.545315	1.460498	0.1346032
5	0.9	0.230746	0.227524	9.23774	3.8538	0.385522	8.876794	2.732862	0.3576490
10	0.4	0.006120	0.001749	9.87653	0.5695	0.000028	9.865037	0.477537	0.0001713
10	0.5	0.007774	0.003455	9.82228	0.7829	0.002989	9.812958	0.618592	0.0015107
10	0.6	0.016354	0.012161	9.72373	1.1271	0.015115	9.716133	0.852531	0.0092000
10	0.7	0.040675	0.036607	9.54612	1.6814	0.056359	9.525983	1.265490	0.0433613
10	0.8	0.095757	0.091942	9.15435	2.5931	0.178238	9.098354	2.055569	0.1681403
10	0.9	0.330487	0.327367	9.00565	4.5981	0.360757	7.918966	3.700704	0.4302934
20	0.4	0.007869	0.002951	9.82219	0.6871	0.000782	9.733936	0.671095	0.0002637
20	0.5	0.012012	0.007207	9.74126	0.9354	0.003211	9.628082	0.872976	0.0022319

- model validation in a test bed
  - generation of HTTP traffic by Apache benchmark tool *ab*
  - only **HTTP 1.0** used and *renewal model* applied
  - model parameters: hmin = 5, hmax=10, N=30, L = 40
- relative accuracy of the model subject to the measurements:
  - mean number of idle processes: 3%; variance 20 % deviation; abserr(wait prob) <8%

Part III

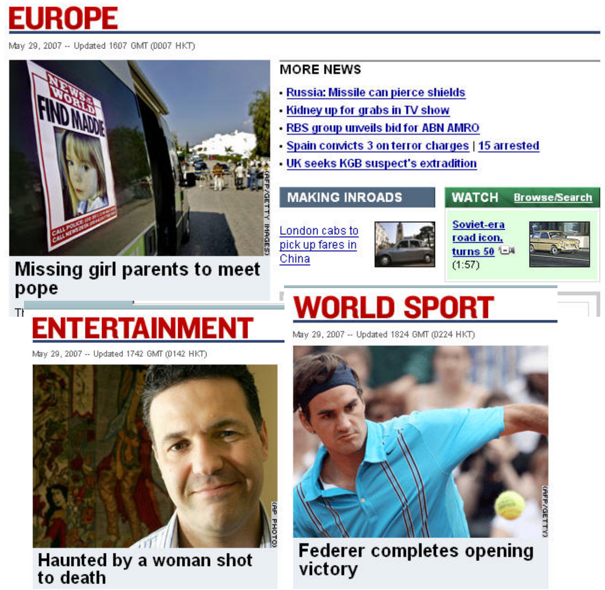
III. Service Analysis of a Media Server

Hierarchical Workload Model



Object-Dependent Arrival Streams

- popularity-dependent request patterns for streamed objects
  - modelled by M object classes i
  - requests determined by Zipf-Mandelbrot law

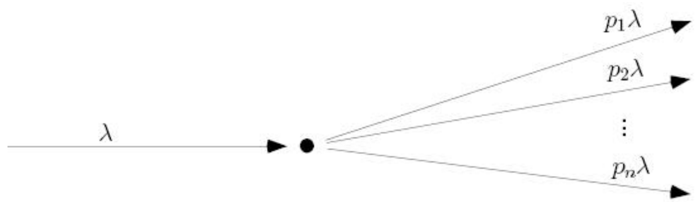


$$f_i(\alpha, k) = C \cdot \frac{1}{(i+k)^\alpha}, \quad i \in \mathbb{N}, k \in \mathbb{N}, \alpha \in \mathbb{R}^+$$

$$C^{-1} = \sum_i \frac{1}{(i+k)^\alpha}$$

  - simplest heavy-tailed case used here: Zipf law

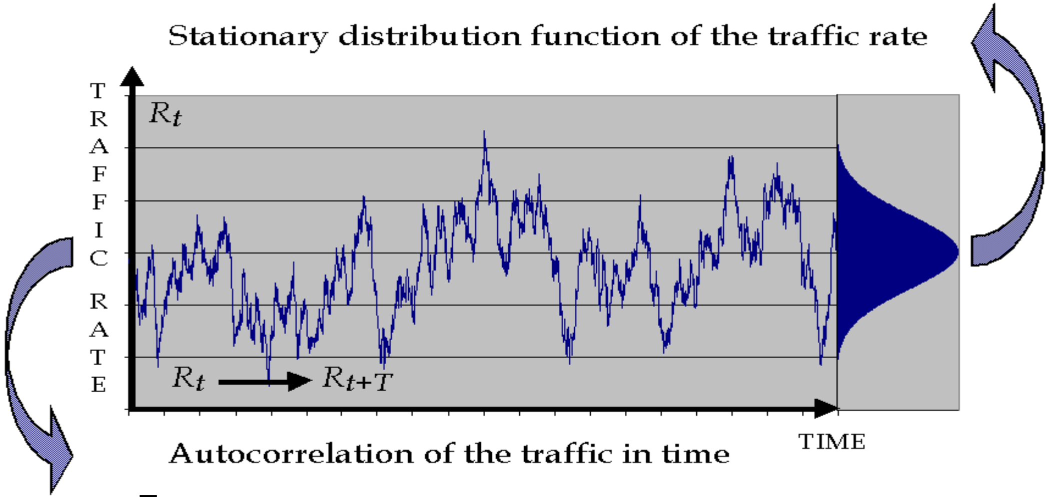
$$f_i = f_i(\alpha, 0) = C \cdot \frac{1}{i^\alpha}$$



- request streams to objects
  - modelled by M Poisson streams of rate λ_i
  - rate estimation using non-homogeneous Poisson processes

Arrival Rate of Requested Streaming Bandwidth

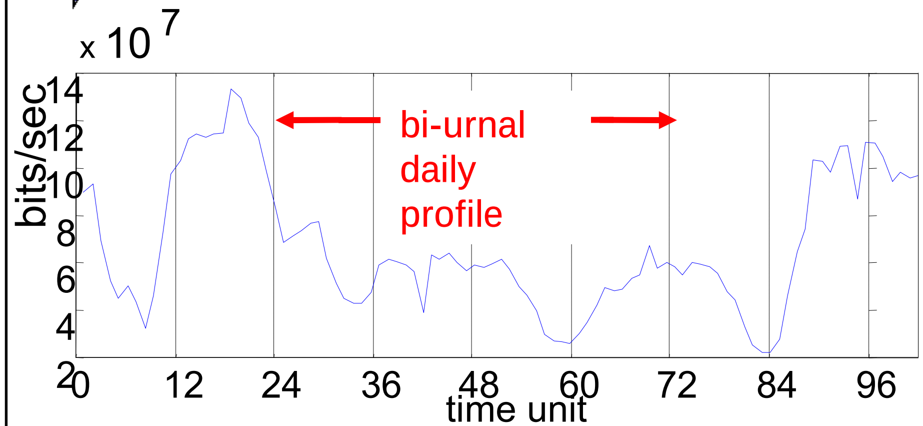
•bandwidth profile



WebTC-Modeling of MPEG-4 stream

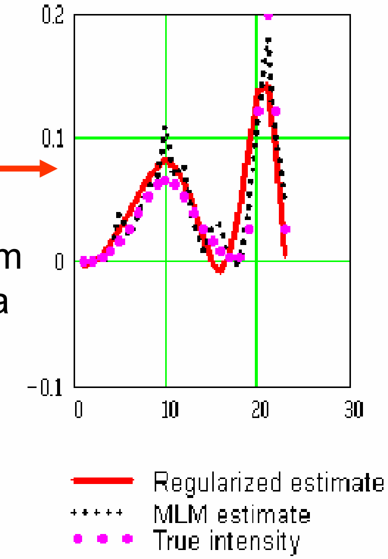
• SARIMA process

• cf. EuroNGI WebTC project, Deliverable, 2007



Method: regularization of a statistically ill-posed problem by a Fourier expansion of a NHPP rate

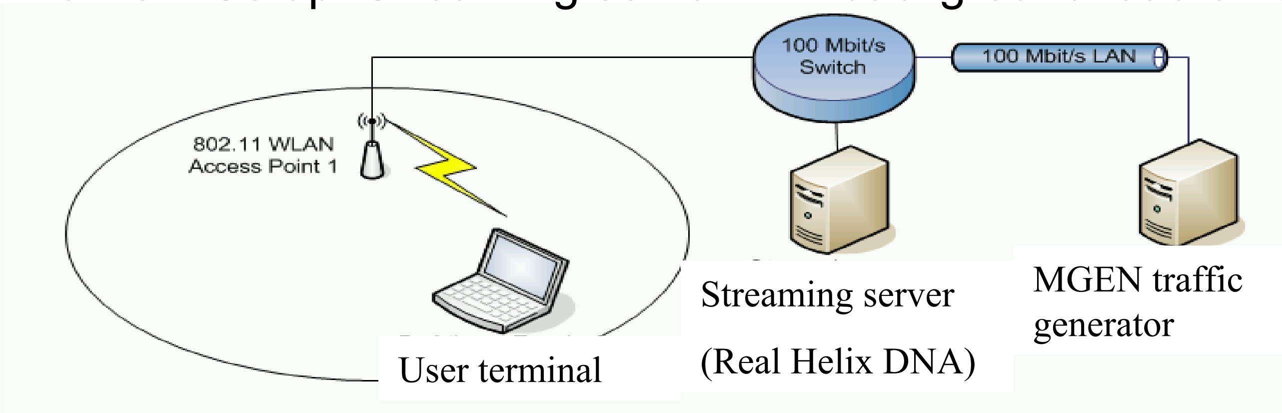
$$\lambda^{\gamma}(t) = \sum_{j=1}^{N} \alpha_j^{\gamma} \varphi_j(t)$$



• apply similar approach to NHPP estimation of session arrival rates

Wireless Testbed with Interfering Traffic

Network Setup: Streaming server with background load on an 802.3 LAN

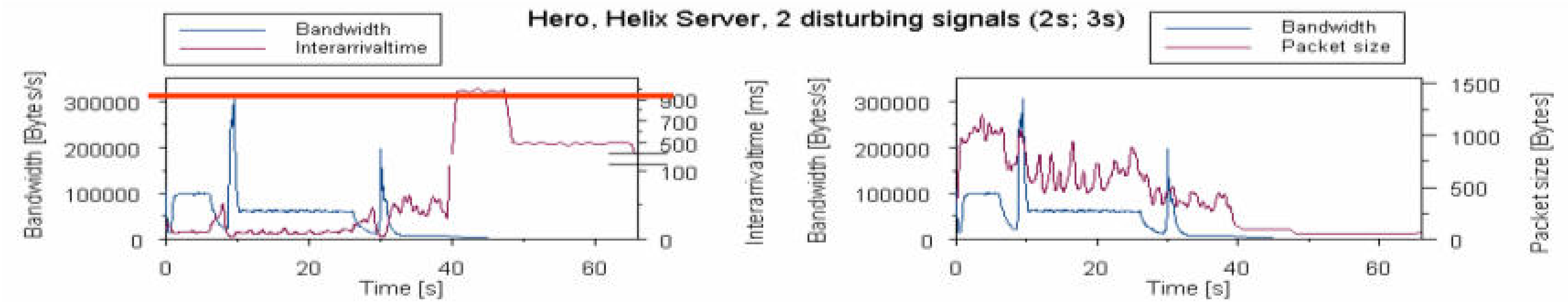


- Other source:

HP study of streaming server request behavior, Computer Networks, 2007

- line of sight connection between AP and a portable terminal
  - no movement of the terminal
- disturbing background traffic on the link
  - various traffic patterns generated on the wireless link

Reactive bandwidth requirement of a video streaming system (Real's Helix server)



Bandwidth Modeling

- Bandwidth modeling
  - apply basic bandwidth units (BBU) of fixed sized b_0
    - e.g. b_0=64 kbps or 32 kbps
  - determines K *traffic classes*

- Bandwidth requirement
  - depends on used codec and spatial resolution

SQCIF	(128 x 96)
QCIF	(176 x 144)
CIF	(352 x 288)
4CIF	(704 x 576)
16CIF	(1408 x 1152)

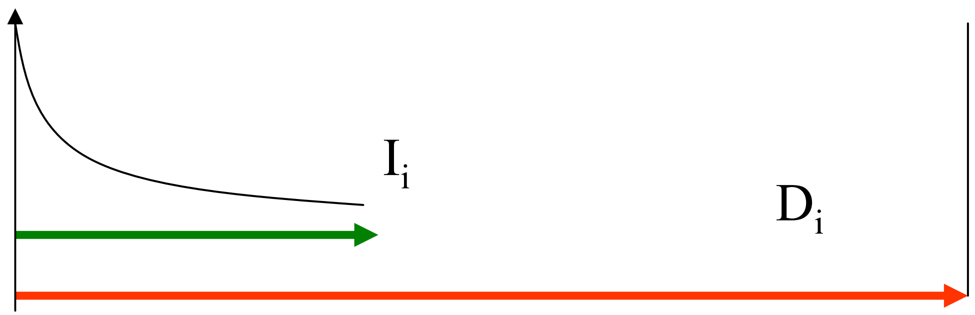
  - can use a *peak-bandwidth* description
    - approach applied here
  - or *sustainable* or *equivalent bandwidth* model (R. Guerin and others)
    - often too complex
    - requires burst level modeling, e.g. by on-off sources, at packet level
  - traffic classes of objects

$$TC_i = \{j \in \{1, 2, \ldots, M\} | b(O_j) = c_i\}$$

Modeling Service Durations

- three main object-dependent components for each class i
  - *playback time* D_i
    - determined by length and rate of the video stream
    - deterministic length of mean $1/d_i$
    - modelled by an exponential distribution
  - *inspection time* I_i
    - determined by object type and user behavior
    - mixture of exponential and uniform distribution
    - modelled by an exponential distribution with mean $1/\chi_i$

  - selection ratio β_i for a full playback of the video stream



Modeling Service Durations

- Service time distribution function

$$
\begin{aligned}
F_i(x) = \mathbb{P}\{S_i \le x\} &= \beta_i \mathbb{P}\{D_i \le x\} + (1-\beta_i)\mathbb{P}\{\min(I_i, D_i) \le x\} \\
&= \beta_i \mathbb{P}\{D_i \le x\} + (1-\beta_i)[1 - (1 - \mathbb{P}\{I_i \le x\})(1 - \mathbb{P}\{D_i \le x\})]
\end{aligned}
$$

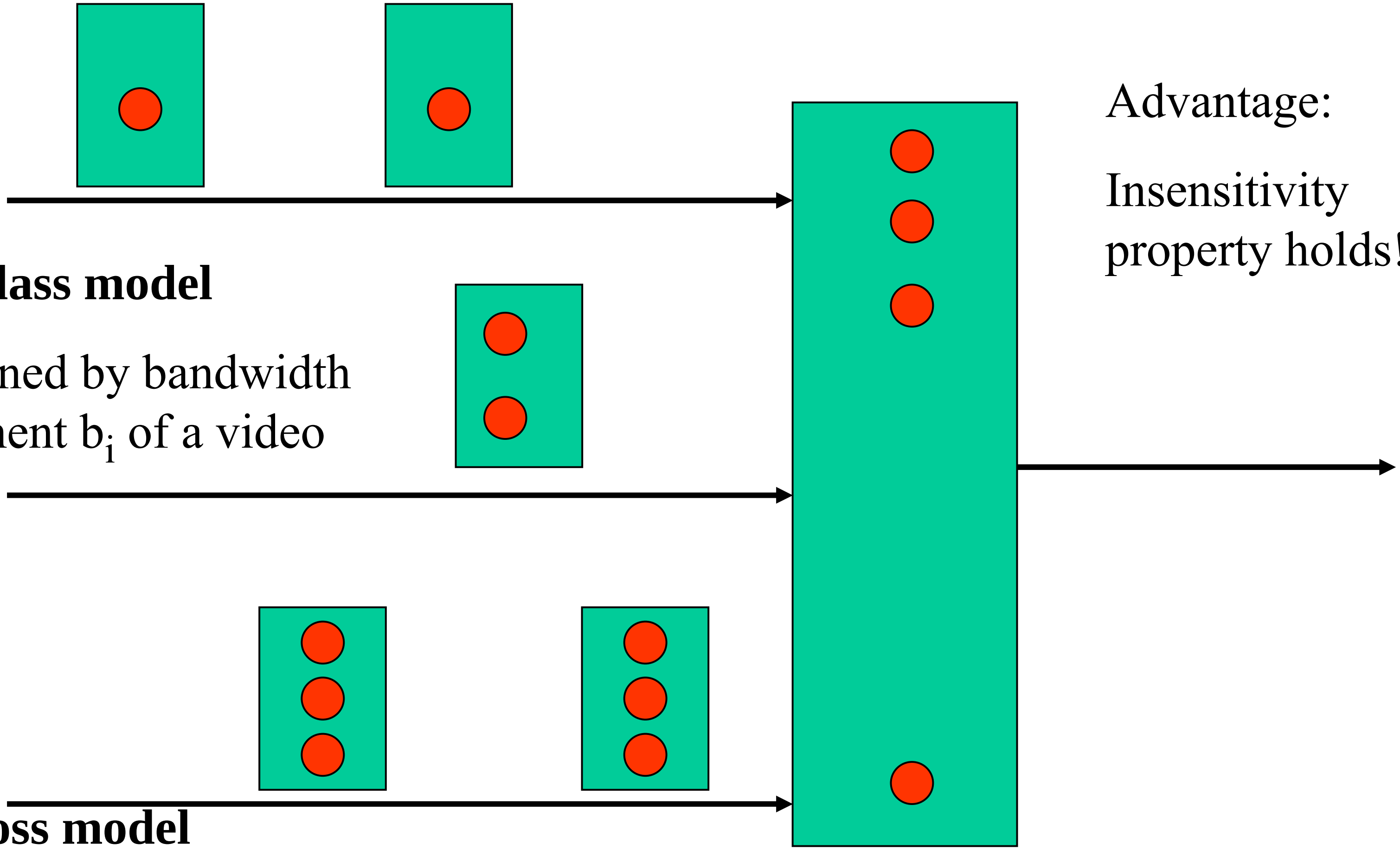
- Mean service time

$$
\begin{aligned}
\mathbb{E}(S_i) &= 1/\mu_i = \beta_i \frac{1}{d_i} + (1-\beta_i)\frac{1}{d_i + \chi_i} \\
&= \frac{\beta_i \chi_i + d_i}{d_i(d_i + \chi_i)} \\
&= \frac{\mathbb{E}(S_i)(\beta_i \mathbb{E}(S_i) + \mathbb{E}(I_i))}{\mathbb{E}(I_i) + \mathbb{E}(S_i)}
\end{aligned}
$$

- Mean service time of an aggregated traffic stream

$$
\mathbb{E}(T_i) = 1/\tau_i = \sum_{j \in TC_i} \frac{\lambda_j}{\xi_i} \frac{\mathbb{E}(D_j)}{\mathbb{E}(D_j) + \mathbb{E}(I_j)} [\beta_j \mathbb{E}(D_j) + \mathbb{E}(I_j)]
$$

Queueing Model of a Streaming Media Server



Advantage:

Insensitivity property holds!

traffic class model

•determined by bandwidth requirement b_i of a video stream

Erlang loss model

• of capacity C BBUs with K different service classes i (e.g. object/traffic classes)

• of arrival rate λ_i, service rate μ_i and bandwidth requirement b_i BBUs

Product-Form Steady-State Distribution

• State space

$$\Sigma = \{n = (n_1, \ldots, n_K) \in \mathbb{N}_0^K \,|\, 0 \le n_i \le \lfloor C/b_i \rfloor, i \in \{1, \ldots, K\}, b^t n \le C\}$$

$$\Sigma(j) = \{n \in \Sigma \,|\, b^t n = j\}$$

$$\Sigma_k = \bigcup_{j=0}^{C} \Sigma(j - b^t e_k) = \{n \in \Sigma \,|\, b^t n \le C - b_k\}$$

$$\Sigma_0 = \bigcup_{j=0}^{C} \Sigma(j) = \{n \in \Sigma \,|\, b^t n \le C\}$$

• Steady-state distribution of class-related Markov vector process $X(t) = (X_i(t))$

$$\pi(n) = \frac{1}{G} \cdot \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!}, \quad n \in \Sigma, \quad G = \sum_{n \in \Sigma} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!} \qquad \rho_i = \rho_i(\alpha) = \lambda_i/\mu_i = \lambda p_i(\alpha)/\mu_i$$

$$q(j) = \sum_{n \in \Sigma(j)} \pi(n), \quad R_k(j) = \sum_{n \in \Sigma(j)} n_k \, \pi(n) \qquad q(j) = \frac{1}{j} \sum_{i=1}^{K} b_i \rho_i \, q(j - b_i), \quad j = 0, 1, \ldots, C$$

$$G(C) = \sum_{n \in \Sigma_0} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!}, \quad G(C - b_k) = \sum_{n \in \Sigma_k} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!}$$

Part III

III.3 Perfomance Analysis of a Streaming Media Server

Performance Characteristics

- Utilization

$$Y = b^t \cdot X,$$

$$U = \mathbb{E}(Y) = \sum_{j=0}^{C} j\mathbb{P}\{b^t X = j\} = \sum_{j=1}^{C} jq(j)$$

- Blocking of service class i

$$B_i = 1 - \mathbb{P}\{b^t X \leq C - b^t e_i\} = 1 - \frac{G(C - b_i)}{G(C)}$$

- Throughput of class i

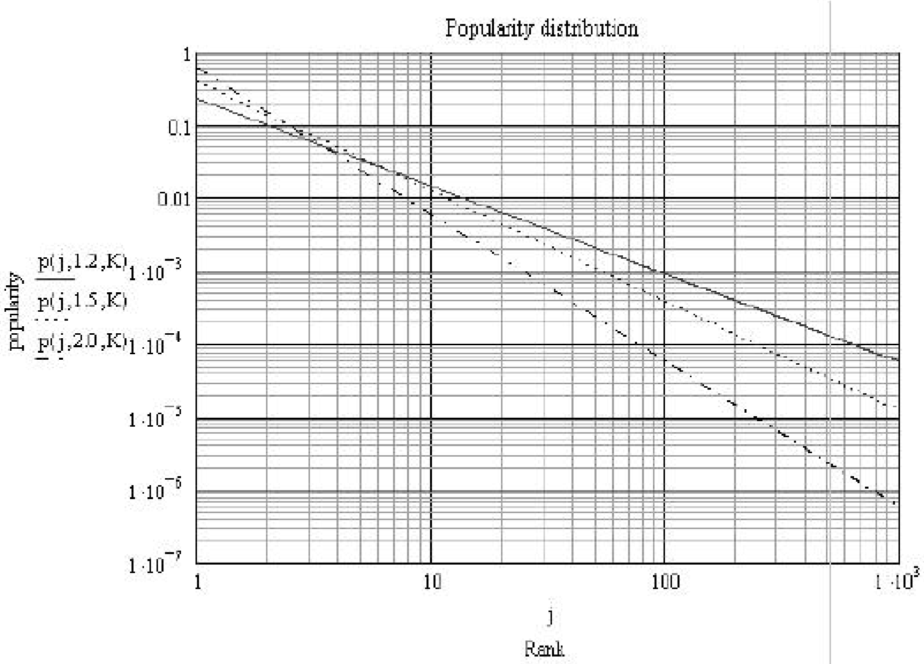
$$T_i = \lambda_i \cdot (1 - B_i)$$

Arrival-Sensitivity on the Tail-Index of the Popularity Distribution

- given Poisson r.v. $Y_j \sim \mathrm{Pois}(\rho_j(\alpha))$

$$\mathbb{P}\{Y_j = n\} = e^{-\rho_j(\alpha)} \frac{\rho_j(\alpha)^n}{n!}$$

- exhibits monotonicity in log-likelihood ordering



- monotone increasing in λ_j decreasing in μ_j
- behavior in α depends on class j
- occupancy X_j and thoughput behavior inherited

Case Study and its Parameter Settings

- Illustration of the Markovian modeling approach

$$
\begin{aligned}
p(1.2,3) &= (0.587, 0.256, 0.157)^t \\
p(1.5,3) &= (0.647, 0.229, 0.124)^t \\
p(2,3) &= (0.735, 0.184, 0.082)^t
\end{aligned}
$$

$$\mu_1 = 1.5, \mu_2 = 1, \mu_3 = 0.833,$$

$$b = (1, 2, 3)^t.$$

Table 1: Study of a streaming media system with three service classes

C	B_1	B_2	B_3	GB	TH_1	TH_2	TH_3	U	U/C
400	0.034	0.067	0.099	0.053	5.672	2.385	1.416	373.497	0.934
300	0.044	0.087	0.128	0.068	5.611	2.334	1.37	279.887	0.933
200	0.064	0.124	0.18	0.098	5.497	2.239	1.288	186.29	0.931
140	0.087	0.166	0.24	0.131	5.363	2.131	1.195	130.146	0.93
120	0.098	0.188	0.269	0.148	5.294	2.076	1.149	111.437	0.929
100	0.114	0.216	0.307	0.17	5.203	2.005	1.09	92.733	0.927
80	0.135	0.254	0.357	0.2	5.077	1.908	1.011	74.037	0.925
60	0.167	0.308	0.427	0.244	4.89	1.768	0.901	55.353	0.923
40	0.22	0.394	0.532	0.314	4.581	1.548	0.735	36.696	0.917
20	0.328	0.554	0.708	0.446	3.946	1.139	0.458	18.103	0.905

Part IV

IV. Analysis of Packet-Switched Multimedia Streams

Part IV

IV.1 Characterization of Multimedia Packet Streams by Measured Data

Operational Modeling and Nonparametric Traffic Estimation

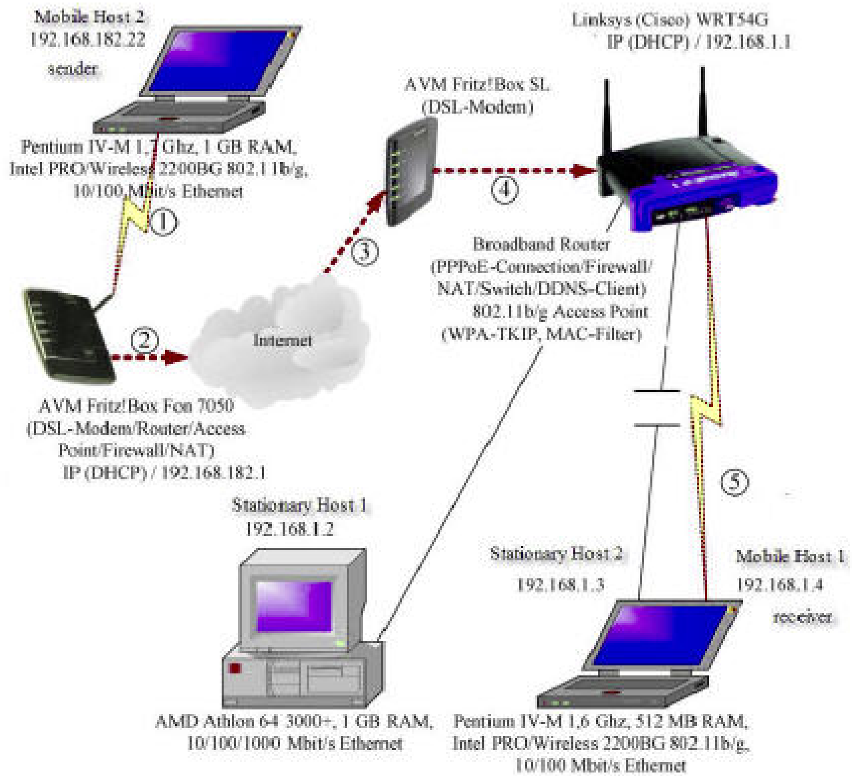
Objectives of the measurement approach

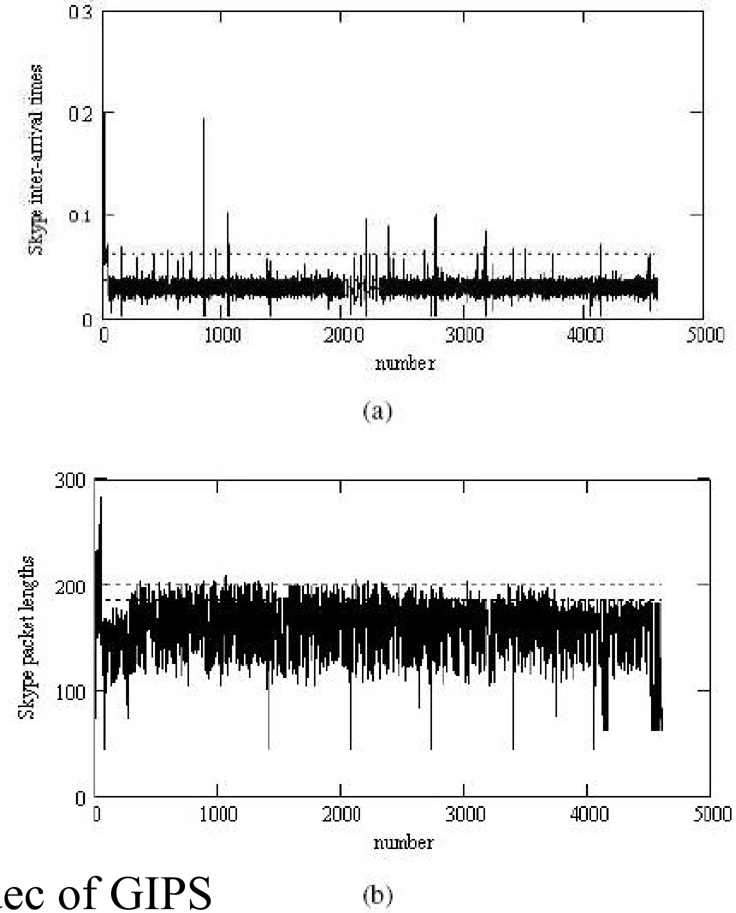
- using *passive measurements* $Z=\{Z_1, Z_2, .., Z_n\}$ at the IP packet layer (or Ethernet layer by monitoring tools like Wireshark, DAG etc.)
- to characterize selected aspects of the transport performance (QoS metrics) and user satisfaction (QoE metrics)
- in terms of bivariate sequences $Z_i = (X_i, Y_i)$ determined by
  - *inter-arrival times* (IATs) X_i between packets of a VBR stream
  - the associated *packet lengths* (PLs) Y_i

- flow characterization on a *virtual* network link by
  - the mean delivery time variation of successfully transferred packets
  - the overall and mean byte loss determined by a bufferless fluid model
  - the lossless periods on a capacity-constrained link
    - motivated by wireless networking over error-prone channels

Skype Traffic Measurements

•prototypical home environment with 2 wireless segments and ADSL attachments





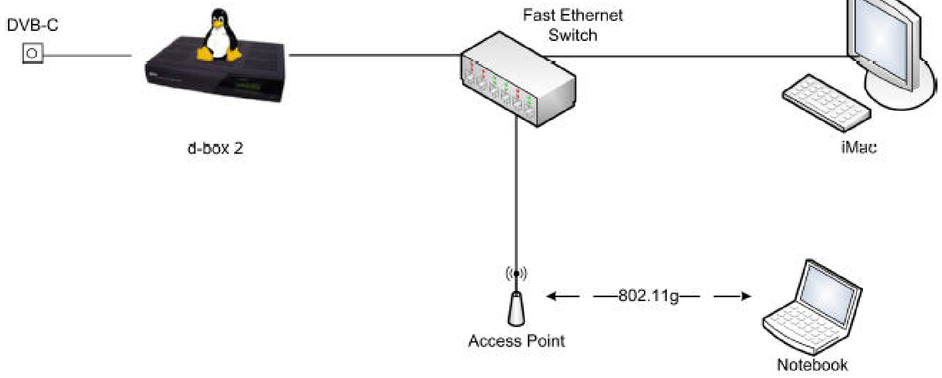
• strongly correlated bivariate time series $Z_i = (X_i, Y_i)$

• due to context-aware adaptive VBR iSAC codec of GIPS

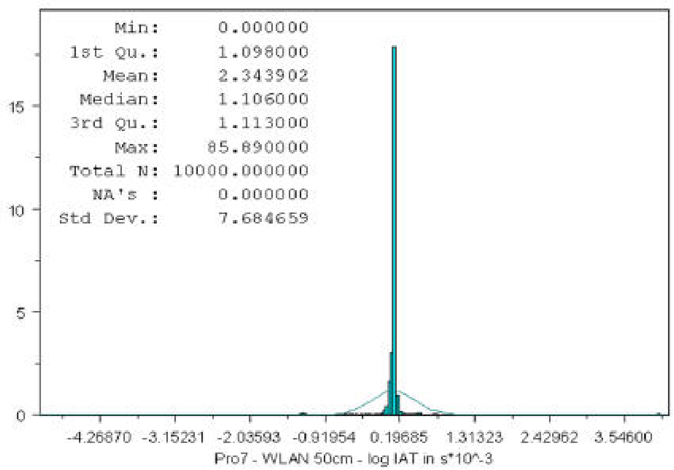
• generating rates in the range of 10 to 32 kbps at a mean IAT of 31 ms & 160 bytes PL

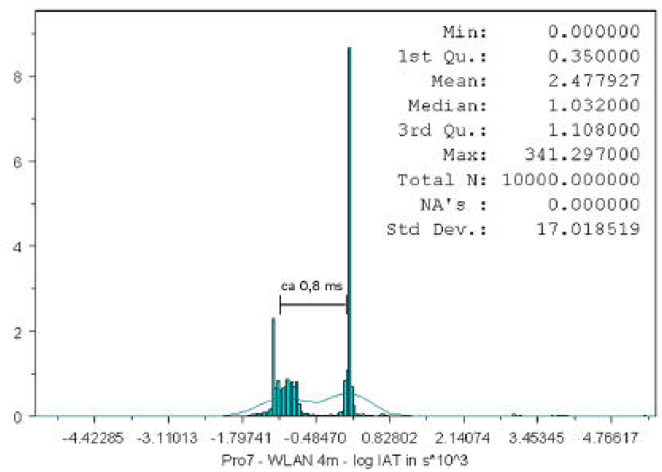
Video Transport of an MPEG-4 Stream in a WLAN

WLAN test bed for the DVB-C transport of an MPEG-4 video streams stemming from a cable network



- inter-arrival times on a log scale
- packet loss causing artefacts





- spreading of IATs span across a wire-less link due to layer 2 retransmis-sion and buffering effects at an AP
- variance increases

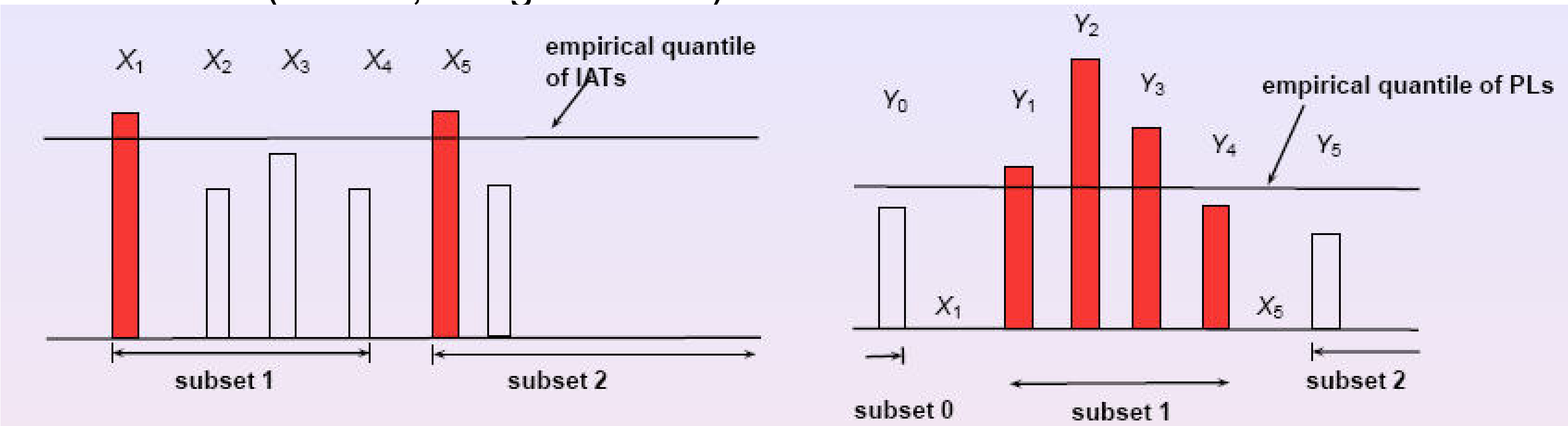
Part IV

IV.2 Statistical Modeling of Qos Indicators by a Bufferless Fluid Approach

Statistical Analysis After Data Partitioning

Requirements of a statistical analysis: stationarity and independence

- pure stationarity not a realistic assumption for real data sets of triple play streams
- independence achievable by partitioning of data into independent blocks of fixed or variable lengths $L_j = \sum_{i=kj}^{kj-1+Nj} X_i$, $j = 1, \ldots, N_s$
  - e.g. by exceedances over the empirical 97% quantiles of the IATs $\{X_i\}$
  - proof of independence by appropriate statistical Portmanteau tests (Runde, Lung-Box etc.)



Features of P2P and VBR Packet Traffic

- often long-range dependent time series of IATs $\{X_i\}$ with heavy-tailed marginal distributions $F(x) = 1 - x^{-\alpha} l(x)$
  - determined by Hill's estimate of the extreme value index $\gamma = 1/\alpha$ on the order statistics $X_{(i)}$

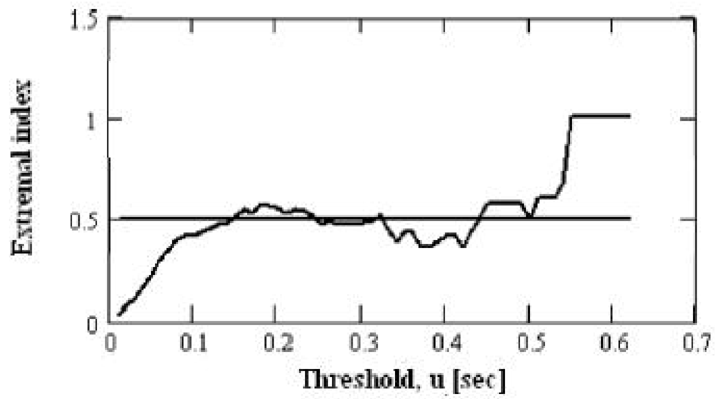
$$\widehat{\gamma}^H(n,k) = \frac{1}{k}\sum_{i=1}^{k} \ln X_{(n-i+1)} - \ln X_{(n-k)}$$

  - extremal index θ of maxima $M_n = \max X_i$ of IATs in blocks

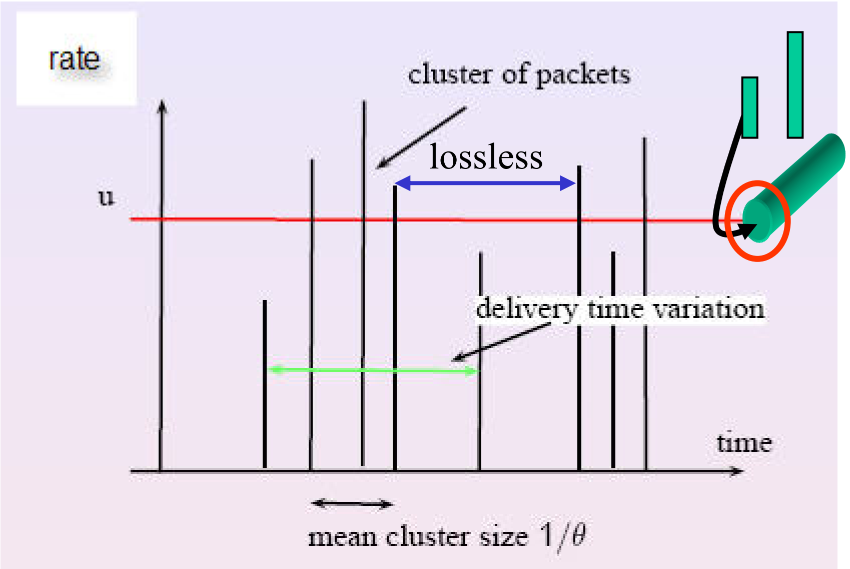
$$\mathbb{P}\{M_n \le u\} \approx \mathbb{P}^\theta\{\widetilde{M}_n \le u\} = F^{n\theta}(u)$$

Features	Skype Data			IPTV Data	
	IAT	PL	Block duration	IAT	IAT Block maxima
Independence	−	−	+	−	+
LRD	+	+	−	+	−
Self-similarity	+	+	+	+	+
Heavy-tailed with finite variance	+	−	−	−	+
Heavy-tailed with infinite variance	−	−	+	+	−
Light-tailed	−	+	−	−	−

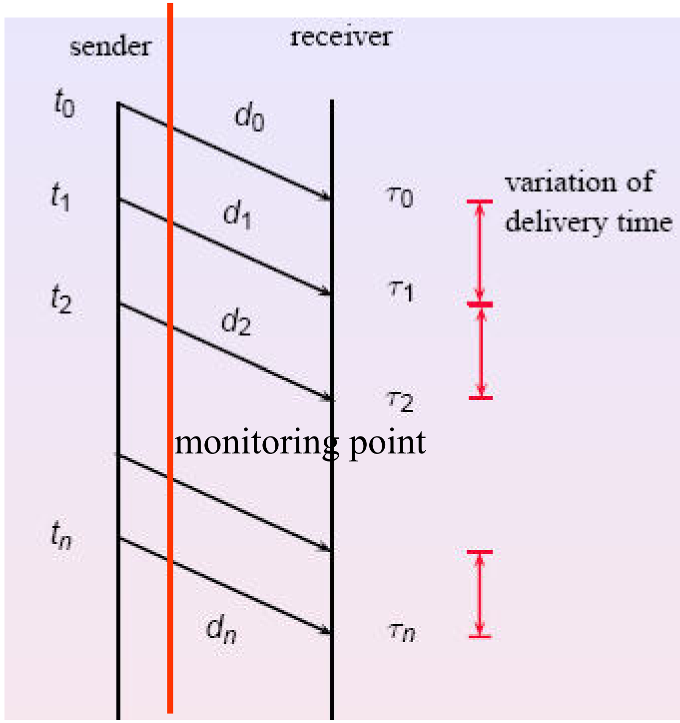
Example: Sopcast over a WLAN



Delivery Time Variation in a Bufferless Fluid Model



- working with representatives of QoS indicators in data blocks (*EVT block method*)
  - such as extreme rate values max R_i
- using instantaneous traffic rates $R_i = Y_i/X_i$ overshooting a capacity threshold u



- determined for completely transferred packets by
  $y_i = \tau_i - \tau_{i-1} = t_i + d_i - (t_{i-1} + d_{i-1}) = \Delta t_i + \Delta d_i$
  IATs at departure Δt_i with mean EX and delay jitter Δd_i

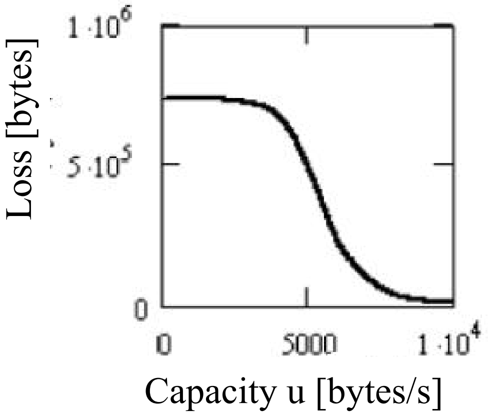
- mean delivery time variation in case of data loss
  $d = (1 + 1/\theta)$ EX
  - with mean cluster size $1/\theta$
- derived from extremal index $\theta \in [0, 1]$
  - as indicator of changes in limiting df of $\max_i R_i$

Part IV

IV.3 Capacity Requirements of VBR and Peer-to-Peer Packet Traffic

Capacity Requirement Subject to Loss Constraints

- Loss determined by estimates of the rate exceedances over a capacity threshold u
  - overall byte loss $$\hat{E}(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(R_i > u)$$
  - mean byte loss $$\hat{e}(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(R_i > u) / \sum_{i=1}^{n} \mathbb{1}(R_i > u)$$
- Example: Skype data with mean E(X) = 31 [ms] and rate estimate θ_R=0.38



- 3% overall byte loss at c = 8.534 kbps
- mean byte loss $\hat{e}(c)$= 168 bytes
- quantiles of lossless period in this case

quantile	75	80	85	95	97
loss-free [ms]	34	35	36	39	44

Part V

V. Conclusions and Open Issues

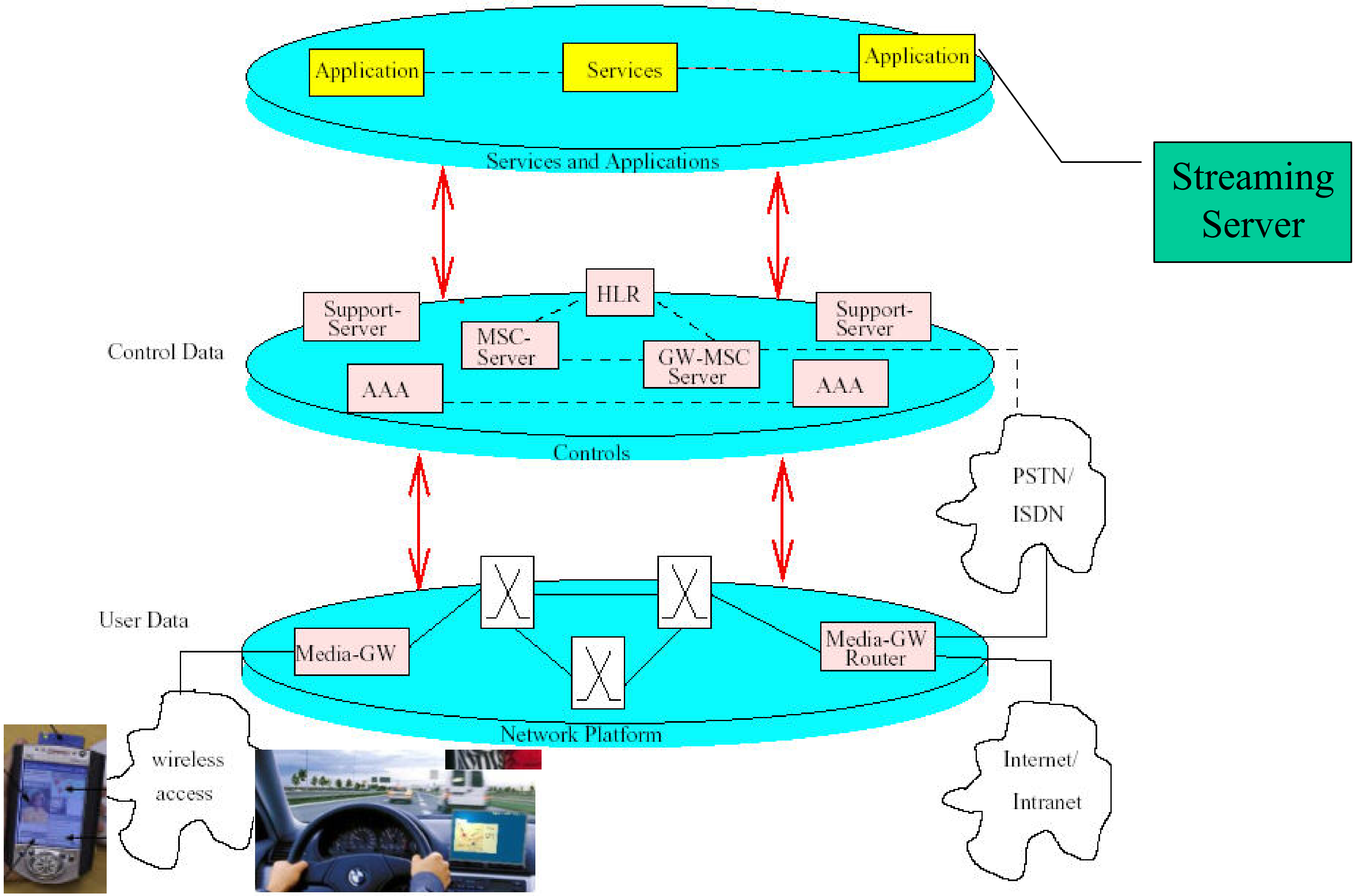
Conclusions

- single-tier Web server architecture
  - study of Apache´s non-threaded multi-processing module Prefork
    - Markovian modeling by a queueing network with batch Markovian arrival process and a dynamic number of servers
  - potential extensions
    - hybrid thread/process UNIX module MPM Worker
    - extension of the QNW model to heavy-tailed service times

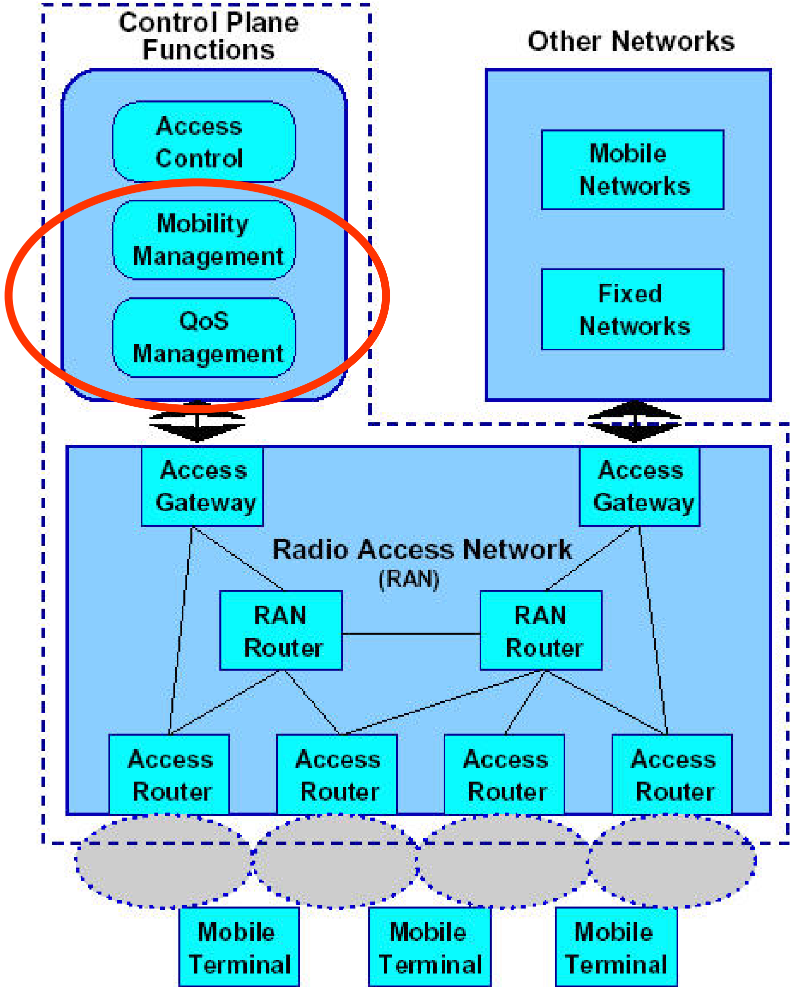
- single-tier streaming media server architecture
  - Markovian modeling by a Erlang loss model
    - with popularity-dependent Poissonian batch Markovian arrival process
      - traffic classes determined by CBR bandwidth requirements
      - service times including object-dependent inspection behavior
    - closed-form analysis of resulting $\Sigma\ M^X/M/C/C$ loss model
  - potential extensions
    - QNW model with heavy-tailed service times
    - optimized access control and bandwidth assignment for a server farm model

- nonparametric statistical traffic characterization and Qos/QoE analysis of VBR or P2P packet traffic

Challenge 1: Impact of Real Mobility in Wireless NGNs



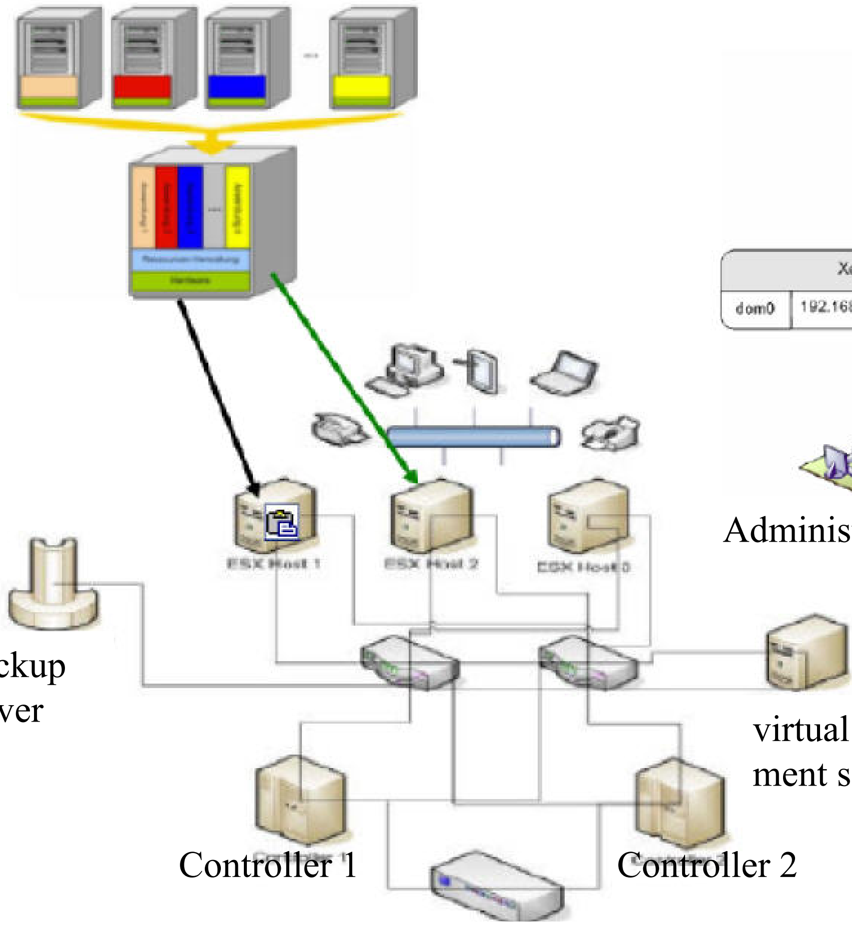
Challenge 2: Interworking of Wireless Network Components

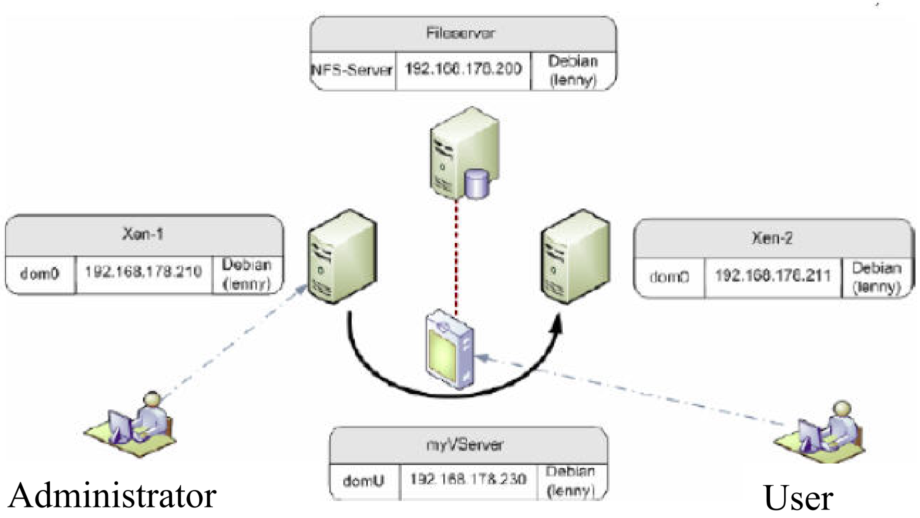


Components of a mobile network system.

Challenge 3: Impact of High Availability Concepts

- Live migration of a virtualized infrastructure supported by heartbeat signaling



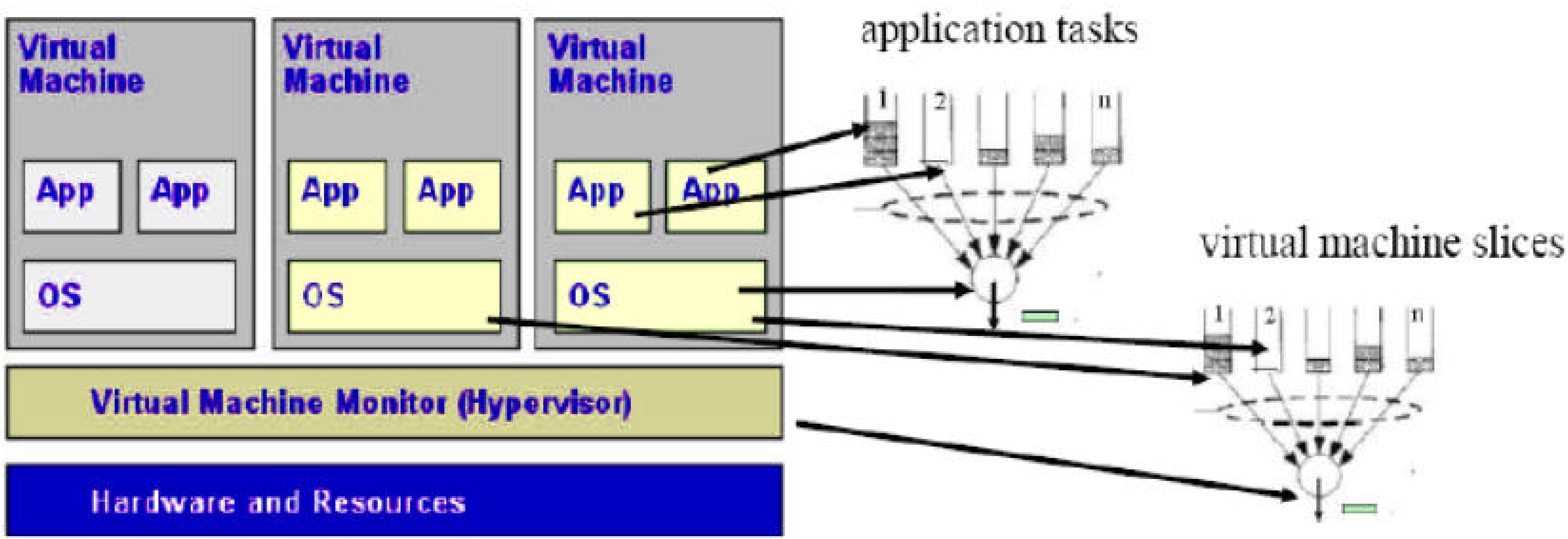


Examples:

- XEN
- VMware Motion
- IETF reliabe server pooling

Challenge 4: Impact of Virtualization on Performance

- impact of hierarchical scheduling at the different processing layers within a media or web server
  - started modeling at one layer by a polling system with phase-type service times in cooperation with G. Mishkoy, Chisinau



References

- T. van Do, U.R. Krieger, R. Chakka: Performance Modeling of an ApacheWeb Server with a Dynamic Pool of Service Processes. Telecommunication Systems, 39(2-3), 117–129, November 2008.
- U.R. Krieger, T. van Do: Performance Modeling of a Streaming Media Server. Proc. Fourth Euro-FGI Workshop on "New Trends in Modelling, Quantitative Methods and Measurements", Ghent, Belgium, May 31 - June 1, 2007.
- Natalia M. Markovitch, Udo R. Krieger: Statistical Analysis of VoIP Flows Generated by Skype Users, *Proc. IEEE International Work-shop on Traffic Management and Traffic Engineering for the Future Internet*, FITraMEn, Porto, Portugal, 11–12 December, 2008.
- N.M. Markovich, U.R. Krieger: Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic. In D. Kouvatsos, ed., Performance Handbook - Next Generation Internet, Lecture Notes in Computer Science, LNCS 5233, Springer, Heidelberg, to appear 2010.
- U. Krieger et al.: Characterization of Web Services with Advanced Multimedia Components, W.JRA.S.20, Deliverable D.JRA.S.20.1, February 2007, EuroNGI-project: http://*www.eurongi.org*

Acknowledgment

The authors acknowledge the partial support by the following projects:

- IST-FP6-NoE EuroNGI/EuroFGI,
- COST IC0703,
- BMBF-project MDA 08/15,
- a Werner von Siemens Excellence Award by Siemens Hungary.