MAPS: System Support for Mobility and Adaptation in Pervasive Services

Cheng-Zhong Xu
Dept of Electrical & Computer Engg.
Wayne State University, Detroit, Michigan
http://www.ece.eng.wayne.edu/~czxu
czxu@wayne.edu

March, 2004
Pervasive Internet Services

• **New information services**
  - News, stock, weather, etc
  - Location-aware: ATM, restaurant, parking
  - M-commerce: banking, ticketing, etc

• **New communication services**
  - Email, Chat, Instant Message
  - Voice, Telephony, Video conf.

• **Services accessible anytime and anywhere**
Characteristics

• Diversity
  - Diverse Access Networks:
    • PSTN, Bluetooth, Cellular, DSL, Cable, LAN, Satellite, etc
  - Diverse Access Devices
    • PDA, phone, computer, “Dick Tracy” watch, etc

• Resource-constrained
  - Info processing capacity: cpu, memory
  - Storage, networking,
  - Battery power, etc

• Mobile
  - Mobility is an inherent nature of human being, moving toward resource or away from scarcity.
  - User (device) and computation
Design Goals

• Scalable and Secure Service Arch.
  - Rapid development/deployment of new services

• Mobility Support: access on-the-move
  - User/Device vs Computation

• Adaptation: proactive in response to change
  - user requirements, preferences,
  - available resources and operation conditions

• Acknowledgements:
  - NSF and NASA
Presentation Outline

• Continuous media adaptation for service differentiation (IEEE Trans. on PDS)

• Mobility support for network-centric, composable services (J. Concurrency, IEEE MDC'03)

• Semantic prefetching to reduce Internet access latency (IEEE Trans. on KDE)
Service Differentiation

- Provide different QoS levels to different requests
  - SD vs Best-effort, same service to all
- Perspectives at network edge
  - On an indiscriminate Web site (IEEE IPDPS’04)
    - Control behaviors of aggressive clients for fairness
  - On a Web content hosting site
    - Treat clients of different content providers differently
  - On an E-commerce site (IEEE ICDCS’04)
    - Give higher priority to sessions of buyers than visitors
  - On a streaming site (IEEE Trans. on PDS)
    - Adapt to various devices & access patterns
S.D. on Streaming Servers

- **QoS Metrics**
  - Bandwidth: frame rate, encoding bit rate, etc
  - Latency: queueing delay

- **Resource-constraints**
  - Net-IO bandwidth
  - Storage

Transcoding-enabled SD
Problem Statement

• When, and how to schedule streaming requests so as
to provide predictable and controllable fair-sharing
(PCF) services
  - Predictability: schedules must be consistent, independent
    of variations of the class workloads
  - Controllability: controllable parameters to adjust quality
    factors between classes
  - Fairness: lower classes not be over-compromised, especially
    when workload is high
Related Work

- Recent content adaptation work mostly focused on making a trade-off between content quality and response time [Almeida99] [Fox98] [Chandra00],
- But, they provided no PCF guarantees
- **Our approach: Harmonic BW allocation**
  - Maximize bw utilization
  - Meet PCF requirements
Bandwidth Allocation Model

• Channel as streaming bit rate for minimum QoS of a stream; used as bw allocation unit.
  - Arrival rate of requests of each class $\lambda_i$
  - Channel allocation rate $\mu_i$

• Quality factor of request class $i$:
  $$q_i = \frac{\mu_i}{\lambda_i}$$

• Resource constraint (B is number of available channels)
  $$\sum_{i=1}^{n} \mu_i \leq B$$

• Service availability: (K is the highest encoding bit rate)
  $$1 \leq q_i \leq K$$
Alg 1: Proportional-Share Alloc

- Given quality diff. weights $\delta_i$ and $\delta_j$ of class-i and class-j, allocate channels so that
  \[
  \frac{q_i - 1}{q_j - 1} = \frac{\delta_i}{\delta_j}
  \]

- PS allocation rate:
  \[
  \mu^*_i = \lambda_i + \frac{\lambda_i \delta_i}{\sum_{i=1}^{n} \lambda_i \delta_i} (B - \lambda)
  \]

- PS allocation guarantees PCF schedules for requests of different classes on streaming servers.
- But, it may not be optimal with respect to an overall channel utilization.
Alg. 2: Harmonic BW Alloc

• Given quality diff. weights $\delta_i$ and $\delta_j$ of class-i and class-j, allocate channels

Minimize $\sum_{i=1}^{N} \delta_i \cdot \frac{1}{q_i-1}$

Subject to

1. $\sum_{i=1}^{n} \mu_i \leq B$
2. $1 \leq q_i \leq K$

• Optimal channel allocation rate

$$\mu_i^* = \lambda_i + (B - \lambda) \frac{\sqrt{\delta_i \lambda_i}}{\sum_{i=1}^{n} \sqrt{\delta_i \lambda_i}}$$
Properties of Harmonic Alloc

Harmonic Quality Factor:

\[ q_i^* = 1 + (B - \lambda) \frac{\sqrt{\delta_i \lambda_i}}{\lambda_i \sum_{i=1}^{n} \sqrt{\delta_i \lambda_i}} \]

Properties

- Diff. weight of a class increases, its quality factor increases at the cost of others.

- Quality factor of a class drops with the increase of its arrival rate

- Load variation of a higher-weighted class causes a bigger quality change of others.

- Any load shift from a higher-weighted class to a lower-weighted class leads to an increase of the quality of each class.
Proportional-sharing Property

- The optimal harmonic allocation guarantees a proportional share distribution of excess bandwidth

\[ \frac{q_i^* - 1}{q_j^* - 1} = \sqrt{\frac{\lambda_j}{\lambda_i}} \sqrt{\frac{\delta_i}{\delta_j}} \]
Simulation Model

Scheduling process

Bandwidth release rate

video duration

current period

Request arrival rate

\[ \lambda_1 \]
\[ \lambda_2 \]
\[ \vdots \]
\[ \lambda_n \]

smoothing window

Feedback queue 1,2,...,N

admission control

A(10%) B (40%) C(50%)

4 : 2 : 1
Impact of SD on rejection rate
Impact of SD on queueing delay
Proportional BW Allocation
Summary

- Harmonic bw allocation for service differentiation on streaming servers
  - Transcoding-enabled dynamic content
  - Maximize the bw utility
  - Meet PCF requirements
Presentation Outline

• Continuous media adaptation for service differentiation

• Mobility support for network-centric, composable services

• Semantic prefetching to reduce Internet access latency
Mobile Agent Computing

- **Mobility**
  - Nature of human being, moving toward resources or away from scarcity
  - User/device mobility vs computation migration

- **Mobile agent:**
  - Autonomous agent, acting behalf of its client
  - Able to travel from machine to machine on open and distributed systems, carrying its code, data, and running state.
Why Mobile Agent Computing?

• Execute remotely, asynchronously and autonomously
  - Disconnection is no longer exception

• Reduce the network load and overcome network latency
  - Move composable service to data, instead of passing data to service

• Adapt to the change of environment
  - Extra measure for eternal services: service never shutdown

• Encapsulate protocols (self explained data)
  - Automate deployment and upgrade
  - Extra security measure
Challenges

• Middleware support for transparent migration
  - What to do, how to do, and where to do
  - Itinerary as an additional integral component

• Security is a primary concern on open env.
  - Protect hosts from malicious agents
  - Protect agents from untrustful hosts
Mobile Agent Infrastructure

- **Script language based**
  - Telescript [White 94], Agent Tcl [Gray 96]
  - Tacoma [Johansen 98]

- **Java based**
  - Aglet [Lange'98], Concordia[Wong'97], Voyager[Objectspace97], etc.
  - **Naplet** at WSU, starting from 1998. Latest release Naplet 0.18
    - Experimental systems in support of composable adaptive services
    - Open source code: [http://www.cic.eng.wayne.edu](http://www.cic.eng.wayne.edu)
    - Flexible: Separation of policy from mechanism
Naplet Features

- Structured navigation mechanism
  - Seq, Par, Alt is regular-completed
- Connection migration with mobile agents
- Proportional-share resource management
- Agent-oriented access control
  - Delegation
  - Temporal/Spatial
MAIL: Mobile Agent Itinerary Lang.

Itinerary $i = \text{skip} | \text{stop} | \text{visit} | i_1 ; i_2 | \text{if } c \text{ then } i_1 \text{ else } i_2 | i_1 \parallel i_2 | \text{while } c \text{ do } i$

Regular-completeness:
- Any regular trace can be expressed by MAIL.
- An example of non-regular trace:
  Site a being visited $x$ times followed by visits of b for the same number of times
Access Control in Role-based Agent System

Agent-oriented access control

- Privilege delegation to agent
- Agent authentication
- Agent permissions, constraints

Spatial-Temporal constraints

- Unified time frame
- Temporal permission and validity duration obligation
- Access history and spatial satisfaction obligation
Apps I: Network Management

MAP: Mobile Agent Producer
CNMP: Conventional Net Management Protocol
MAEE: Mobile Agent Execution Environment
SNMP: Simple Network Management Protocol
MIB: Management Information Base
Experimental Results

SNMP vs Bcast Naplet

Naplet vs Multithreading
Apps II: Mobile Grid Computing

- Service composition
  - Composable service
  - Components are agile
- Service brokerage
  - Resource supplier
  - Resource demand
- Adaptive to change of environment
- Fault tolerant

- J. of Concurrency (2000)
Mobile Virtual Machine

- Once a server becomes unavailable, migrate running agents to new machine
- Upon a request for more resources, clone agents on additional machines

Hybrid mobility
- Service migration
- Agent migration
- Interplay of service and agent migration
Experimental Results

Cost for forming an agent-based virtual machine

LU Factorization of 256 x 256 double matrix
Migration Decision for Hybrid Mobility

Load balancing in heterogeneous reconfigurable VMs

Hybrid mobility
  • Service migration
  • Agent migration

Migration decision optimization
  • Optimal migration timing
  • Lower bound of destination server capacity
  • Interplay of service & agent migration
Summary

- Mobility support for composable services
  - Naplet mobile agent infrastructure
  - MAIL itinerary language
  - Agent-oriented access control and Temporal/Spatial constraints
  - Migration decision for hybrid mobility
  - Mobile code applications
Presentation Outline

• Continuous media adaptation for service differentiation

• Mobility support for network-centric, composable services

• Semantic prefetching to reduce Internet access latency
Personalized Info. Assistant

• **User intention capture**
  - what kinds of news?
  - shopping favorites?
  - Edit a document ~ printing service
  - Out-of-town travel ~ food service
  - etc

• **Prediction of future action**
  - Prefetching to alleviate access latency
  - Pre-processing for service composition, adaptation
PIA for Web Prefetching

- Personalized info. agent watches over surfing
- Extract semantic knowledge about the activities
- Prediction
- Prefetching and caching
An Example: NewsAgent

- Based on keywords in anchor text (news caption);
- News category to resolve keyword polysemy;
- Keyword weight determines current surfing interest
- Neural network to adjust the weights in response to the change of surfing interest
Evaluation Methodology

- Set up as a client-side browser proxy
- Cross-examination in daily surfing of news sites: abcnews, msnbc, cnn
- Using one site for training and the other two for testing of prediction accuracy
- The experiment was conducted when “Clinton” affair and “Kosovo” event were widely reported.
Training Process

Access pattern in training

Success rate in training
Evaluation Results

Hit ratio and byte hit ratio

Waste ratio and byte ratio
Conclusions

• **MAPS**: system support for mobility and adaptation in pervasive services
  - Modeling, analysis, and prototyping of several new technologies:
  - Prefetching, content adaptation, service differentiation, mobile agent, load balancing, connection migration, etc.

• What’s next?
  - Each technology needs to be further studied
  - More importantly, robust integration of the technologies into Internet services.