CS 414 – Multimedia Systems Design
Lecture 29 – Media Server (Part 4)

Klara Nahrstedt

Spring 2009

Administrative

- **MP4 is out** – deadline Friday, May 1, 2009 – final competition (5-7pm), in 216 SC
  - Pre-competition of all projects, Thursday, April 30, 5-7pm in 216 SC
- **Discussion session** – Tuesday, April 14, 7pm in 3401 SC

Outline

- **Problem:**
  - VOD service offers a large selection of videos from which customers can choose – want to offer low access latency for customers
- **Main Challenge:**
  - How to handle large number of customers, maintain low cost of operation and at the same time provide acceptable access latency
- **Caching**
  - Source: Caching Techniques for Streaming Multimedia over the Internet, Markus Hofmann, Eugene Ng, Katherine Guo, Sanjoy Paul, Hui Zhang
- **Batching**
  - Source: Selecting among Replicated Batching VOD Servers, Meng Guo, Mustafa Ammar, E. Zegura
- **Patching**
  - Source: Hierarchical Video Patching with Optimal Server Bandwidth, H. Hlavacs, S. Buchinger

True Video-On Demand System

- **True VOD:** serve thousands of clients simultaneously and allowing service any time (variable access time)
- **Goal:** minimize the required resource consumption such as
  - **Server bandwidth** (disk I/O and network) – amount of data per time unit sent from server to clients
  - **Client bandwidth** – network bandwidth that a client must be able to receive
  - **Client buffer requirements** – amount of data client has to be able to temporarily store locally
  - **Start-up delay** – time between issuing request for playback and start of playback

VOD System Delivery Schemes (to handle large number of clients)

- **Periodic broadcast**
  - Data-centered approach
  - Server channel is dedicated to video objects (movie channel) and broadcasting periodically
- **Scheduled multicast**
  - User-centered approach
  - Server dedicates channels to individual users
  - When server channel is free, the server selects a batch of clients to multicast according to some scheduling policy
- **Server replication**
  - Servers maintaining the same videos are placed in multiple locations in the network
  - Server selection is a main issue

Caching for Streaming Media

- Caching – common technique to enhance scalability of general information dissemination
- Existing caching schemes are not designed for and do not take advantage of streaming characteristics
- Need New Caching for Streaming Media

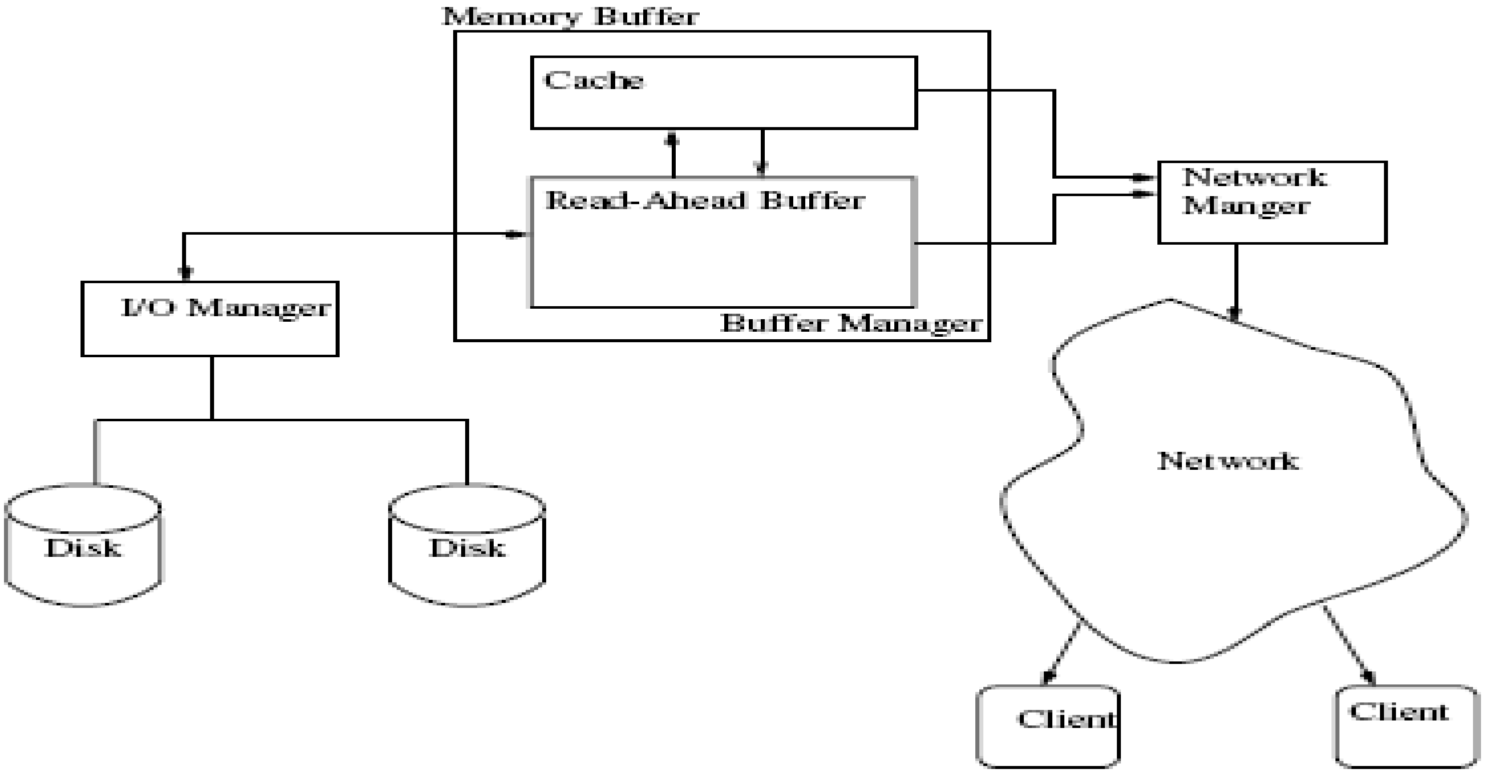
Techniques for Increasing Server Capacity

- Caching
  - Interval Caching
  - Frequency Caching
- Key Point
  - In conventional systems, caching used to improve program performance
  - In video servers, caching is used to increase server capacity

Caching

- Read-ahead buffering
  - Blocks are read and buffered ahead of time they are needed
  - Early systems assumed separate buffers for each clients
- Recent systems assume a global buffer cache, where cached data is shared among all clients

Caching in Media Servers



Source: “preemptive, but safe interval caching for real-time multimedia systems “Lee et al. 2003

Interval Caching

- This caching exploits sequential nature of multimedia accesses
  - Two streams Si and Sj are defined as consecutive if Si is the stream that next reads data blocks that have just been read by Sj. Such a pair of consecutive streams are referred to as preceding stream and following stream.
- Interval caching scheme exploits temporal locality accessing the same MM object, by caching intervals between successive streams (preceding stream and following stream)
  - The interval caching policy orders all consecutive pairs in terms of increasing memory requirements.
  - It then allocates memory to as many of consecutive pairs as possible

Interval Caching

- Memory requirements of intervals are proportional to length of interval and play-out rate of streams involved
- When interval is cached, following stream does not have to go to disk, since all necessary data are in cache
- Algorithm:
  - Order intervals based on increasing space –smaller interval implies smaller time to reaccess
  - Optimal for homogeneous clients
- Dynamically adapts to changes in workload

Frequency Caching

- Typical video accesses follow **80-20 rule** (i.e., 80% of requests access 20% of video objects)
- Cache most frequently accessed video objects
- Requires large buffer space
- Not dynamic
  - frequency determination is based on past history or future estimates/Zipf distribution

Taxonomy of Cache Replacement Policies

- Recency of access: locality of reference
- Frequency based: hot sets with independent accesses
- Optimal: knowledge of the time of next access
- Size-based: different size objects
- Miss cost based: different times to fetch objects
- Resource-based: resource usage of different object classes

Patching

- **Stream tapping or patching** – technique to support true VOD;
- Patching assumes **multicast transmission** and clients arriving **late** to miss the start of main transmission
- These late clients immediately **receive main transmission** and store it temporarily in a buffer.
- In parallel, each client connects to server via **unicast** and **transports (patches)** the missing video start which can be shown immediately

Types of Patching

- Greedy Patching
  - a new main transmission is started only if the old one has reached the end of the video
  - Clients arriving in between create only patching transmissions
- Grace Patching
  - If a new client arrives and the ongoing main transmission is at least **T seconds old**, then the server automatically starts a new main transmission which plays whole video from the start again.

Types of Patching

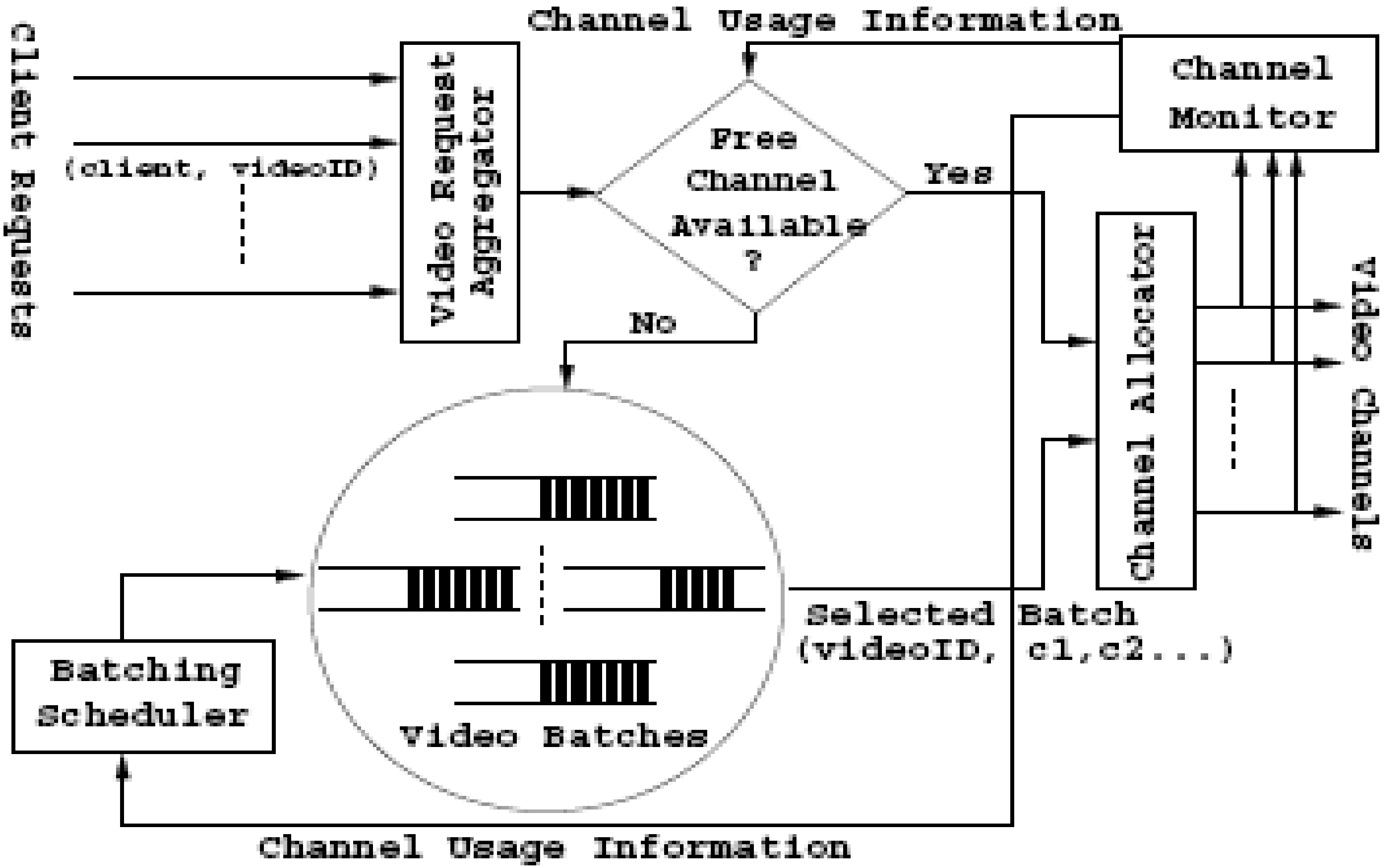
- **Two-Level Patching**
  - Clients **share main transmission** as well as **patches**
  - Main transmission runs on **level zero** and patches patching the start of the main transmission run on **one-level** patching channels.
  - This approach depends on
    - Time for sharing a one-level patch
    - T - Number of periods until zero-level new main transmission is shared

Batching

- **Batching** – grouping clients requesting the same video object that arrive within a short duration of time or through adaptive piggy-backing
- Increasing batching window increases the number of clients being served simultaneously, but also increases reneging probability
  - reneging time – amount of time after which client leaves VOD service without delivery of video
  - *Increasing minimum wait time increases client reneging*
- **Performance metrics**: latency, reneging probability and fairness
- Policies:
  - FCFS, MQL (Maximum Queue Length), FCFS-n

Batching Policies

- FCFS: schedules the batch whose first client comes earliest, with the aim of achieving some level of fairness
- Maximum Queue Length: schedules the batch with largest batch size, with the aim of maximizing throughput
- FCFS-n: schedule the playback of *n* most popular videos at predefined regular intervals and service the remaining in FCFS order



Source: “Selecting among Replicated Batching VOD Servers, Guo et al. 2002

Conclusion

- Designers of VOD systems strive to achieve low access latency for customers
- Challenges:
  - Handle large amount of customers (clients)
  - Maintain low cost of operation
  - Provide acceptable access latency
- Caching, Patching, Batching are examples of techniques to achieve these goals