

# Research Plan for PhD Degree

Topic: Adaptive Content Distribution in the Future Internet

Student: Ossi Karkulahti (Department of Computer Science at University of Helsinki)

Supervisor: Jussi Kangasharju (Department of Computer Science at University of Helsinki)

Starting date: Fall 2009

Projected time of dissertation: End of year 2013

## 1 Adaptive Content Distribution in the Future Internet

Currently the Internet is full of variable content such as web sites, photos, movies and songs, which all pose different challenges for efficient distribution. Until recently most of the content was provided by large institutions and commercial organisations. However, in recent years user-generated content has grown drastically and sites like YouTube [ ] and Wikipedia [31] are hugely popular nowadays. An important part of this research is to determine appropriate content distribution mechanisms for different kinds of content providers.

We define end-users as such who want to consume the content that is provided by the above mentioned content providers. In the scope of this research we consider end-users to be actual humans, although we do acknowledge the fact that machine-to-machine communication is likely to grow in the future. However, our stance is that a system designed according the constraints set by a human user, is likely to work for machine-to-machine information distribution as well.

End-users access the content produced by the content providers through the Internet. The internet, in turn, consists of many independent network providers. Usually, end-users requests for content and content providers replies with content will travel through a number of network providers.

Today the content is distributed using either web servers, content distribution networks (CDN), or peer-to-peer networks. Majority of the content is so-called bulk data,

that are web-pages, software or media downloads. Streaming content corresponds only to a small fraction of the overall content volume. Thus, the emphasis of this research is on the delivery of the bulk content.

Related work includes previous efforts made in the area of web caches [5, 7, 11, 32], containing research made with replacement policies [32] and hierarchical organisation of the caches [5, 25, 28] in order to improve the overall system performance. CDNs are covered in the work described in [1, 10]. A lot of progress have been made in peer-to-peer content distribution, including work done to evaluate its effectiveness [2] and the development of new algorithms and mechanisms [3, 12, 18, 19, 20, 21].

There has been fairly small amount of research on evaluating the effectiveness of the different content distribution algorithms listed above. In the context of web caching, the main focus has been on looking at cache replacement algorithms [9, 15, 32], cache placement [22, 24], and linking different caches to build larger caching systems [14, 26, 29, 30]. Also, work evaluating the effects of parallel downloads has been performed [27].

In the context of CDNs, less work has been done, possibly partly due to the proprietary nature of commercial CDNs. Existing work includes evaluation of CDN redirection algorithms [16], placement of content [17], as well as development of new CDN systems [10].

Evaluating peer-to-peer content distribution has not yet been done on large scale. The existing work includes [8, 13], which both come to the same conclusion that peer-to-peer-based content distribution can have significant benefits if it is done correctly, meaning, if the content is retrieved from nearby peers instead of random peers which typically happens in BitTorrent-like systems. However, [8] is based on a simulation study and [13] only looks at the network effects and not how users would experience the downloads.

The goal of this research is to evaluate how the different kinds of information can be delivered in an efficient manner. We emphasis on gaining understanding about fundamental differences between different content distribution systems. Furthermore, we will try to develop algorithms, mechanisms, and architectures that focus on efficient

content distribution on the Internet. The difference to existing solutions is that our objective is to produce comprehensive and adaptive content distribution algorithms and mechanisms, that are able to adapt their behaviour according to changing conditions, such as changes in network topology and content popularity. In addition, this research is likely to be the first of its kind to address the problem of user-generated content distribution, as so far no evaluation about how users access the user-generated content and how it should be distributed has been conducted.

The research will consist of three main phases.

- Evaluation of content distribution architectures from the points of view of content providers, network operators, and users.
- Measurements of actual systems, with focus on measuring peer-to-peer systems from the user's point of view, and on measuring systems with user-generated content.
- Development of new algorithms and mechanisms and their evaluation.

## 1.1 Research Questions, Objectives and Methods

Table 1 lists the most notable of our research questions and Table 2 has the primary objectives of our research.

We will tackle the first research question in the evaluation phase of our research. That will complete our first objective and partly the second one. The evaluation is performed analytically and by using simulation and real-world data about Internet topology. The simulations are conducted using tools such as PlanetSim [23] and ChunkSim [6].

1	What are the fundamental differences between different content distribution systems?
2	How, when, and why users generate content?
3	How, when, and why end-users access content?
4	How do different kinds of content providers distribute different content types efficiently?
5	What kind of algorithms, mechanisms, and architectures are needed for efficient and adaptable content distribution?
6	How to build adaptive content distribution system atop of the current network and physical layers?
7	What are the difference in implementing adaptive content distribution in the realms of mobile and wired networks?

*Table 1: Round-up of Research Questions*

In measurement phase we will answer research questions numbers 2, 3, 4, and 5. Furthermore, the through the second phase we will complete our objectives 2, 3, and 4. The measurement phase has two parts. In the first one we will measure BitTorrent's [4] behaviour from user's point of view, with aim to collect up-to-date data about how content is being distributed in today's Internet. The data will allow us to determine network cost of distribution and, also, estimate download times accurately. The second part of the measurement phase focuses on analysing how user-generated content is accessed and distributed. We will use, for example, popular photo sharing site [www.flickr.com](http://www.flickr.com) to gather data for the analysis process.

In the development phase we will use the knowledge acquired in the previous phases and give answers to the research questions 6 and 7, and achieve the remaining objectives 4 and 5. Development will carried by simply creating scripts and extension modules that run atop of the current network and physical layers. Currently there is no plan to create a full implementation.

1	Gain understanding about the fundamental differences between distribution systems
2	Evaluate how the different kinds of information can be delivered in an efficient manner
3	Determine which distribution architecture is most suited for which kind of content
4	Develop algorithms, mechanisms, and architectures for efficient content distribution
5	Provide comprehensive and adaptive content algorithms and mechanisms

*Table 2: Objectives of Research*