

# Identifying file-sharing P2P traffic based on traffic characteristics.

Valerij Trajt

18. Januar 2010

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Analyse von Traffic-Charakteristiken BitTorrent-ähnlicher Anwendungen</b>	<b>2</b>
2.1	Definition von TCP/UDP-Flows . . . . .	2
2.2	Analysemethode . . . . .	2
2.3	Traffic-Charakteristiken von BitTorrent . . . . .	2
2.4	Traffic-Charakteristiken anderer Anwendungen . . . . .	4
<b>3</b>	<b>RHD und RPD-basierte Identifikation von Klasse 1-P2P-Traffic</b>	<b>6</b>
3.1	Definition von RHD . . . . .	6
3.2	Vergleich des RHD von Traffic verschiedener Anwendungen . .	6
3.3	Definition von RPD . . . . .	7
3.4	Algorithmus . . . . .	8
<b>4</b>	<b>Fazit</b>	<b>10</b>

# 1 Einleitung

Inzwischen gibt es viele verschiedene Peer-To-Peer-Systeme (P2P-Systeme). Solche Systeme werden dazu genutzt, die Einschränkungen von Client-Server-Systemen zu überwinden. P2P-Systeme besitzen die Eigenschaft, dass je mehr Nutzer gleiche Datei herunterladen oder gleiche TV-Sendung per P2P anschauen, desto höher die Download-Geschwindigkeit sein wird. Da sich die Nutzer Teile der Dateien und Streaming-Daten gegenseitig zur Verfügung stellen, müssen nur relativ wenige Nutzer auf die eigentlichen Server zugreifen. Als Folge daraus, dass P2P-Systeme mehr Bandbreite als übrige Anwendungen für sich beanspruchen und nutzen, kann es dazu kommen, dass die Performanz der anderen Anwendungen (wie z.B. E-Mail oder Web) beeinträchtigt wird. Deshalb kann die Einschränkung der Nutzung von P2P-Anwendungen während bestimmter Zeiten im Interesse einiger Internet-Service-Provider (ISP), Unternehmen und anderer Interessenten . Um dies zu erreichen, muss zuerst Traffic dieser Art identifiziert werden. Man kann alle P2P-Anwendungen in zwei Klassen unterteilen. Anwendungen, die so viel wie möglich Verbindungen herstellen, werden als Anwendungen der Klasse 1 bezeichnet und solche, die nur relativ wenige Verbindungen zu anderen Servern herstellen, werden als Klasse 2 bezeichnet. Zu den Anwendungen der Klasse 1 gehören z.B. BitTorrent,  $\mu$ Torrent und weitere File-Sharing-Anwendungen. ICQ, Skype und MSN gehören zu den Anwendungen der Klasse 2. Es wäre zu kostspielig P2P-Traffic global einzuschränken, deshalb kann dieser Traffic an den Grenzknoten einzelner lokalen Netzwerke (Router, die ein lokales Netzwerk mit dem Internet verbinden) mit akzeptablen Kosten identifiziert und beschränkt werden. Im weiteren wird nur auf die Identifizierung von Klasse 1-P2P-Traffic generierenden Rechnern an den Grenzen des Netzwerks eingegangen.

Bisher wurden nur Inhalt- und auf bestimmten Port-Nummern oder Anwendungssignaturen basierte Ansätze vorgeschlagen. Solche Ansätze haben einen großen Nachteil, da sie nach jeder Änderung der P2P-Anwendungen oder des P2P-Protokolls aktualisiert werden müssen. Außerdem wird verschlüsselter P2P-Traffic nicht erkannt. Der neu vorgeschlagene Ansatz ist frei von diesen Nachteilen. P2P-Traffic wird anhand von zwei Metriken identifiziert: discreteness of remote hosts (RHD) und discreteness of remote ports (RPD). (RHD: Unbeständigkeit der Anzahl an Servern, zu denen der Host verbunden ist. RPD: Unbeständigkeit der Anzahl an aktiven Ports).

## **2 Analyse von Traffic-Charakteristiken BitTorrent-ähnlicher Anwendungen**

### **2.1 Definition von TCP/UDP-Flows**

Analyse von Traffic-Charakteristiken und Identifikation von Klasse 1-Traffic benötigt die Klassifikation einzelner Datenpakete zu Flows. Ein Flow ist ein 5-Tupel, der aus local IP, local port, remote IP, remote port, protocol besteht. Ein Flow wird als expired (nicht mehr gültig oder inaktiv) bezeichnet, wenn es eine bestimmte Zeit lang keine zu dem Flow gehörenden Pakete beobachtet werden. Datenpakete des Internetprotokolls, die zwischen bestimmtem Remote-Server und bestimmtem lokalen Rechner pendeln, gleichzeitig Transport-Level-Protocol-Data-Units beinhalten und nur begrenzte "Lebensdauer" haben, sollen entweder zum TCP- oder UDP-Flow gehören. Ein IP-Paket, das zu keinem aktiven Flow gehört, resultiert in einem neuen Flow. Ein Flow kann zwei Zustände annehmen: S\_ACTIVE und S\_TIMEOUT. Ein neuer Flow befindet sich im immer Zustand S\_ACTIVE. Gibt es bestimmte Zeit lang keine neuen Pakete mehr, wird der Flow in den Zustand S\_TIMEOUT versetzt. Zwei Flows, die sich gleichzeitig im S\_ACTIVE-Zustand befinden, werden konkurrierende Flows genannt.

### **2.2 Analysemethode**

Um die Traffic-Charakteristiken von Anwendungen zu analysieren, wurde ein spezielles Tool entwickelt, das die Flow-Statistiken aufzeichnet. Es kann IP-Traffic beobachten, IP-Pakete zu Flows zuordnen, die Dauer berechnen, inbound/outbound Oktette zählen, den Zustand des Flows kalkulieren, Anzahl von konkurrierenden Flows ermitteln.

### **2.3 Traffic-Charakteristiken von BitTorrent**

BT-Traffic eines einzelnen BT-Klienten beinhaltet sowohl TCP- als auch UDP-Flows. NetworkAddressTranslation-Geräte (NAT-Geräte) werden in der Regel an den Grenzen lokaler Netzwerke platziert. Traffic vom internen Netzwerk-Host beinhaltet TCP-Flows, die aus folgenden Teilen bestehen:

1. P2P-Verbindungen zu entfernten Hosts mit global einzigartigen Adressen
2. TCP-Verbindungen, um Verbindungen zu anderen Peers (BT-Klienten) in entfernten LANs herzustellen

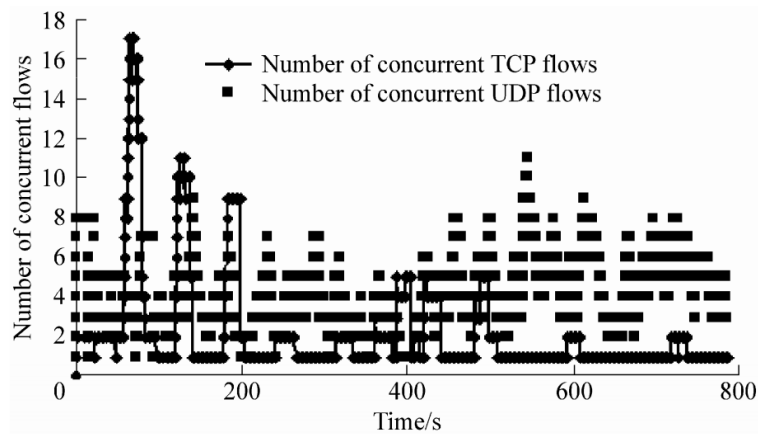
3. TCP-Verbindungen zu Tracker-Servern bestehen. UDP-Flows bestehen aus P2P-Verbindungen zu Hosts in anderen LANs oder zu Hosts, die keine TCP-Verbindungen erlauben, und aus UDP-Flows zwischen dem Host und den Tracker-Servern.

Eigenschaften des BT-Traffics eines Klienten:

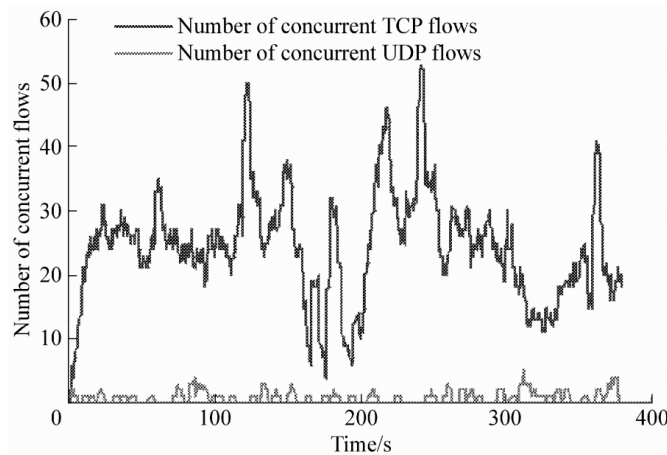
1. TCP- und UDP-Flows koexistieren zwar oft, aber BT-Klient hat selten gleichzeitig TCP- und UDP-Flows von und zu dem Peer-Host.
2. ein BT-Klient hat selten gleichzeitig mehr als eine Assoziierung mit dem gleichen entfernten Host während des Downloads.
3. Remote Endpoints eines BT-Klienten haben in der Regel unterschiedliche IP-Adressen und Portnummern, da jeder BT-Klient praktisch mit beliebigem Port arbeiten kann.
4. Anzahl der konkurrierenden TCP- und UDP-Flows und deren Verhältnis ist nicht stabil. Allgemein hängen die Großen von der Anzahl beteiligter Peers, und von der Art der verwendeten Internet-Verbindungen von Peers ab.
5. Der Anteil von Short-Flows (Flows mit weniger als 5 Paketen) ist relativ hoch. Da einzelne Peers die Verbindung jederzeit unterbrechen können.
6. BT-Traffic beinhaltet mehr Long-Flows als Traffic jeder anderer Art. Ein Long-Flow ist ein Flow, der aus mehr als 100 Paketen besteht und eine Lebensdauer von mehr als 15s. hat.

Um BT-Traffic zu identifizieren, wird die Anzahl von konkurrierenden Flows untersucht.

Erstmals wird eine non-hot-Datei heruntergeladen. Der Traffic wird mit Hilfe des Flows-Statistics-Tool beobachtet. Die Anzahl von konkurrierenden Flows ist grafisch auf dem Bild dargestellt. Die Anzahl von konkurrierenden Flows schwankt mit der Zeit. Die Gesamte Anzahl von konkurrierenden Flows ist relativ klein ( $\leq 10$ ), wenn eine non-hot-Datei heruntergeladen wird. Im Falle einer hot-Datei werden mehr als 10 Flows registriert. Die Übertragungsgeschwindigkeit ist bei einer hot-Datei wesentlich höher. Die Übertragungsgeschwindigkeit hängt von der Anzahl von Peers, zu denen ein BT-Klient verbunden ist (je mehr desto höher).



(a) The number of concurrent flows when downloading a non-hot file (Nov 8 14:08–14:21 2005)

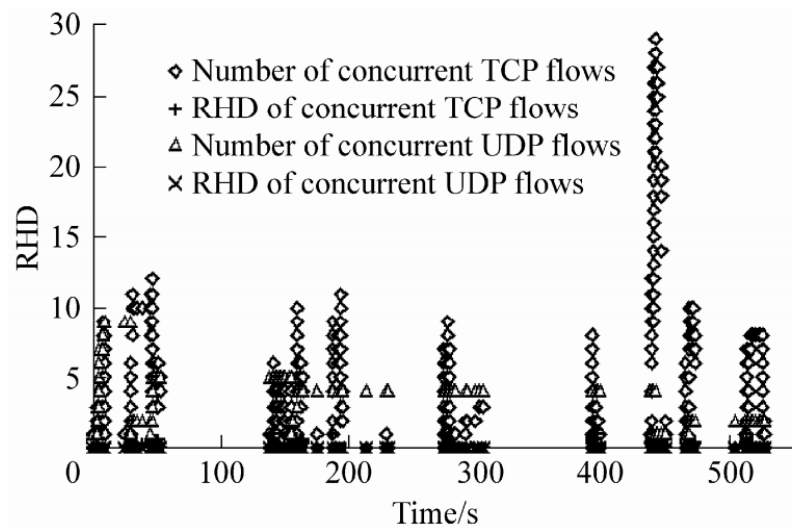


(b) The number of concurrent flows when downloading a fairly hot file (Nov 13 15:02–15:09 2005)

**Bild 1** Anzahl von konkurrierenden TCP/UDP-Flows anhand von der Art der Datei

## 2.4 Traffic-Charakteristiken anderer Anwendungen

Wenn eine Internet-Seite aufgerufen wird, werden alle Elemente zur Darstellung der Seite parallel heruntergeladen, d.h. dass der Web-Traffic eines Benutzers nur wenige konkurrierende TCP-Flows hat, wobei viele Flows den gleichen Endpoint haben können (mehrere Verbindung zum gleichen Server). UDP-Flows in Web-Traffic bestehen nur aus DNS-Anfragen und Antworten.



Ein FTP-Klient stellt immer eine Kontrollverbindung zu dem Server her, um Kommanden zu senden und Status zu empfangen. Falls eine Datei übertragen wird, wird eine neue Verbindung hergestellt. Deshalb gibt's nur 1 oder 2 konkurrierende TCP-Flows und 1 UDP-Flow, wenn man mit einem FTP-Server verbunden ist. Die P2P-Anwendungen der Klasse 2 werden hauptsächlich zur Übertragung von Kurznachrichten (Chat), Audio- oder Video-Streaming genutzt. Die Anzahl von konkurrierenden TCP-Flows hängt von der Anzahl der hergestellten Verbindungen zu anderen Peers. Die Anzahl von konkurrierenden UDP-Flows ist relativ gering.

## 3 RHD und RPD-basierte Identifikation von Klasse 1-P2P-Traffic

### 3.1 Definition von RHD

RHD für konkurrierende Flows wird wie folgt definiert:

$$D_{inst}(t) = \frac{1}{n} \sum_{i=1}^m \log_2 \frac{n}{x_i}$$

mit  $n$  der Anzahl von konkurrierenden Flows im Zeitpunkt  $t$ ,  $m$  der Anzahl von Remote-Netzwerken, zu denen die Flows gehören,  $x_i$  der Anzahl von Flows, deren Endpoints im selben Remote-Netzwerk  $i$  befinden. Die Länge von Netzwerk-Präfix beträgt 23. RHD wird für konkurrierende TCP- und UDP-Flows berechnet.

### 3.2 Vergleich des RHD von Traffic verschiedener Anwendungen

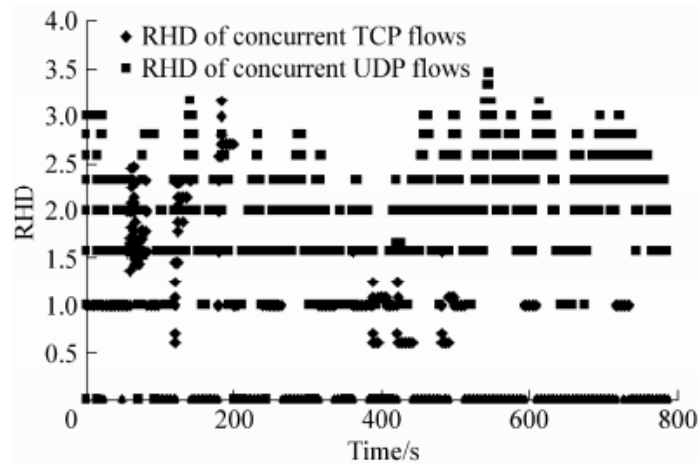
1. RHD von BT-Traffic des einzelnen Host.

Die Anzahl konkurrierender Flows variiert je nach Anzahl und Zustand des Peers, aber mit großer Wahrscheinlichkeit ist der Wert von RHD relativ hoch.

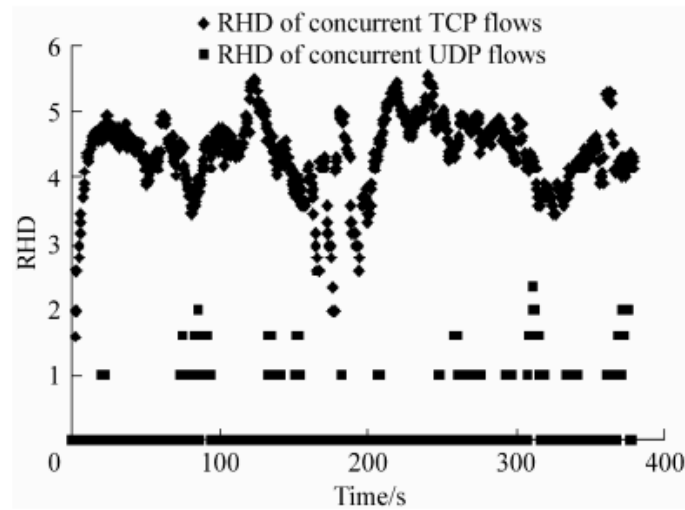
2. RHD von Klasse 2-P2P und anderer Anwendungen.

Wenn man eine Webseite aufruft, könnte die Anzahl von konkurrierenden Flows hoch sein, aber RHD von diesem Traffic ist immer niedrig, denn unterschiedliche Flows haben oft gleichen Remote-Host und Remote-Port. Wenn man Traffic während einer FTP-Session betrachtet, kann man leicht feststellen, dass RHD null beträgt, da der Remote-Host immer gleich bleibt.

Um die Veränderungen der Werte von RHD anschaulich darzustellen, wurden 2 Experimente durchgeführt. Es wurde eine Hot-Datei und eine Non-Hot-Datei heruntergeladen. Die Ergebnisse sehen wie folgt aus:



(a) RHD of concurrent flows when downloading a non-hot file



(b) RHD of concurrent flows when downloading a fairly hot file

### 3.3 Definition von RPD

RPD für konkurrierende Flows wird wie folgt definiert:

$$P_{inst}(t) = \frac{1}{n} \sum_{i=1}^m \log_2 \frac{n}{x_i}$$

mit  $n$  der Anzahl von konkurrierenden Flows im Zeitpunkt  $t$ ,  $m$  der Anzahl von unterschiedlichen verwendeten Ports,  $x_i$  der Anzahl von Flows, deren Remote-Port dem  $i$ . Port gleich ist. RHD wird für konkurrierende TCP- und UDP-Flows berechnet.

### 3.4 Algorithmus

Die Identifikationsmethode basiert auf Analyse der Werte von RHD und RPD des Traffics eines einzelnen Hosts. Diese Methode benötigt die Kontrolle des Traffics jedes einzelnen Hosts des LANs an den Grenzen und benutzt zur Identifikation 2 Kriterien.

**Kriterium 1 (K1)** Momentane RHD-basierte Identifikation: Falls der Wert des momentanen RHD von UDP-Flow des Traffics von Host größer als der Schwellenwert  $D_U$  (z.B. 2.2) ist oder die Summe von den Werten des momentanen RHD von TCP- und UDP-Flows größer als  $D_{SUM}$  (z.B. 2.8) ist, dann kann davon ausgegangen werden, dass ein BT-Traffic vorliegt.

**Kriterium 2 (K2)** RHD-Durchschnitt-basierte Identifikation: Falls die Summe des RHD-Durchschnitts von TCP- und UDP-Flows des Traffics von Host während eines bestimmten Zeitraums  $T$  (z.B.  $T=10$  s) größer als der Schwellenwert  $D_{SumOfAvg}$  ( $\leq D_{SUM}$  z.B. 2.5) ist, dann kann davon ausgegangen werden, dass ein BT-Traffic vorliegt.

Somit sieht der Algorithmus folgendermaßen aus:

I. Kontrolliere jedes eingehende Paket und ordne es dem einzelnen Flow anhand des 5-Tupels und dem Flow-TIMEOUT zu.

II. Gruppiere Flows nach den lokalen IP-Adressen. Ein Flow-Datensatz besteht aus Flow-KEY, FLOW-ZUSTAND, START-TIME (Zeitpunkt der Entstehung), Zeitpunkt des Empfangs des letzten Pakets.

III. Für jeden aktiven Host  $H$  während des gesamten Intervalls der Messung ( $T$  s):

Jede  $G$  Sekunden ( $G < T/20$ ) do:

1. Prüfe, ob jeder aktive TCP- oder UDP-Flow immer noch aktiv ist (Pakete erhält) und versetze den Flow in den Zustand  $S\_TIMEOUT$ , falls nicht.
2. Wenn Anzahl konkurrierender Flows nicht kleiner  $MinPreDispose$  (z.B. 12) ist, dann gruppiere die TCP-Flows anhand des Pseudo-Remote-Endpoints (remote IP, network Prefix, Remote-Port) und sortiere Gruppen mit wenigstens  $MinFCtoWeed$  (z.B. 5) aus. Diese Gruppen werden nicht gelöscht, sondern nur nicht bei der Berechnung von RHD oder RPD berücksichtigt. Mache das gleiche für UDP-Flows.
3. Berechne momentane RHD von TCP- und UDP-Flows. Falls die Anzahl

der konkurrierenden Flows gleich 0 ist, dann setze RHD auf 0.

4. Falls der Wert von RHD zwischen [LowRHD, HighRHD] liegt, dann gruppierere konkurrierende Flows entsprechend ihren Remote-Ports, berechne RPD und setze den Wert des momentanen RHD auf  $w_1 D_{inst}(t) + w_2 P_{inst}(t)$ , (LowRHD = 0.5, HighRHD= 3.5,  $w_1 = 06$ ,  $w_2 = 0, 4$ ).
5. Wende Kriterium 1 an, um zu unterscheiden, ob im gesamten Traffic BT-Traffic enthalten ist. Falls ja, dann gib eine entsprechende Meldung aus.
6. Aktualisiere den RHD-Durchschnitt von TCP- und UDP-Flows. Sei  $D_{inst}^{(k)}$  der Wert des k. momentanen RHDs. Der RHD-Durchschnitt wird wie folgt definiert:  $D_{avg}^{(k)} = \frac{k-1}{h} D_{avg}^{(k-1)} + \frac{D_{inst}^{(k)}}{k}$ , mit  $k \geq 1$

Wende Kriterium 2 an, um zu unterscheiden, ob im gesamten Traffic BT-Traffic enthalten ist. Falls ja, dann gib eine entsprechende Meldung aus.

Lösche alle Aufzeichnungen von den Flows im S\_TIMEOUT Zustand.

Traffic der Klasse 1 kann relativ effizient mit Hilfe des Kriteriums 1 identifiziert werden. Falls der Schwellenwert zu hoch angesetzt wurde, kann ein BT-traffic mit niedriger Anzahl von beteiligten Peers übersehen werden. Demgegenüber wird normaler Traffic als BT-Traffic angesehen, wenn der Schwellenwert zu niedrig ist. Deshalb sollte der Schwellenwert höher angesetzt werden.

## 4 Fazit

Dieser Artikel beschäftigt sich mit der Identifikation von dem P2P-File-Sharing-Traffic eines einzelnen Hosts, die auf Analyse besonderer Charakteristiken basiert. Mit der vorgeschlagenen Methode kann man BT-Traffic in 95% der Fälle richtig erkennen. Diese Methodik benötigt weder Kenntnis von dem P2P-Protokoll noch Packet-Payload-Analyse. Da der Algorithmus auf Charakteristiken basiert, kann BT-Traffic unabhängig davon, ob Verschlüsselung eingesetzt wird oder nicht, identifizieren. Die Tatsache, dass bei der Identifizierung keine Software-Signaturen genutzt werden, ist diese Identifikationsmethodik im Gegensatz zu Kontent-basierten Ansätzen sehr skalierbar.