Problemy wdrażania szerokopasmowych usług multimedialnych w heterogenicznych sieciach IP*

Agnieszka Chodorek (e-mail: a.chodorek@tu.kielce.pl)
Katedra Telekomunikacji, Fotoniki i Nanomateriałów, Politechnika Świętokrzyska

Robert R. Chodorek, Agata Krempa, Andrzej R. Pach
(e-mail: {chodorek, pach}@kt.agh.edu.pl; akrempa@gmail.com)
AGH Akademia Górniczo-Hutnicza w Krakowie, Katedra Telekomunikacji

STRESZCZENIE

Przez ponad 30 lat swojego istnienia, sieć Internet wyewoluowała z klasycznej sieci transmisji danych w największą z sieci wielousługowych. Część usług internetowych tworzona była od podstaw (jak np. poczta elektroniczna), część z nich (np. telewizja, czy telefonia internetowa) stanowi przeniesienie do Internetu usług realizowanych dotychczas w sieciach specjalizowanych (telewizyjnej, telefonicznej).
Wdrażanie nowych usług jest zagadnieniem złożonym. Nowe usługi mogą spowodować problemy z funkcjonowaniem aplikacji już istniejących. Problemy mogą wystąpić także z usługami przeniesionymi z sieci specjalizowanych – w sieci Internet mogą one źle funkcjonować, np. ze względu na dużo niższą jakość transmisji niż wymagana. W niniejszym artykule dokonano analizy współczesnych usług, m.in. z punktu widzenia wymagań, jakie stawiają one dla sieci transmisyjnej. Przeanalizowano najistotniejsze elementy heterogenicznych sieci IP pod kątem współpracy z różnego typu usługami. Określono, jakie problemy występują podczas współistnienia wielu różnych aplikacji w sieci heterogenicznej. Wskazano również, jak te problemy rozwiązywać.

Słowa kluczowe: *sieci IP, usługi multimedialne, protokoły transportowe, zapobieganie i przeciwdziałanie przeciążeniom, sieci heterogeniczne*

ABSTRACT

Problems of implementation of broadband multimedia services in heterogeneous IP networks

For over 30 years of its existence, the Internet has evolved from traditional data networks in the largest multiservice networks. Some Internet service has been created from scratch (like e-mail), some of them (eg. television, or Internet telephony) is the adaptation of services previously implemented in specialized networks (broadcast television network, Plain Old Telephone Service – POTS, etc.).
Implementation of new services is a complex issue. New services may interact with existing applications. Problems can also occur with services adapted from specialized networks - for example, due to much lower transmission quality than required. In this paper an analysis of contemporary services and requirements is described. We show problems, which occur due to the heterogeneity of multiservice network, as well as how to solve these problems.

Key words: *IP networks, multimedia services, transport protocols, congestion control, heterogeneous networks*

1. Wprowadzenie

W ostatniej dekadzie obserwujemy znaczące zmiany w technice i metodach dostępu do sieci Internet. Z punktu widzenia „szarego użytkownika" zmiany te są najbardziej widoczne w przypadku łączy dostępowych, przewodowych i bezprzewodowych. Obejmują one zarówno znaczące zwiększenie dostępnych przepływności, jak i zmianę charakteru połączenia z Internetem. Tam, gdzie jeszcze dziesięć lat temu w powszechnym użyciu były drogie w eksploatacji połączenia wdzwaniane, których koszt zależał od czasu trwania połączenia, obecnie dominują ekonomiczne połączenia stałe, o kosztach zależnych od ilości przesłanych danych.

Drogi dostęp do Internetu wymuszał na użytkowniku pracę w postaci krótkich sesji realizowanych przez modemy korzystające z sieci telefonicznej. Obecnie, przy zastosowaniu innych technologii dostępowych (jak np. ADSL – *Asymmetric Digital Subscriber Line*, czy sieci 3G/4G), sesje uległy wydłużeniu, niekiedy znacznemu. Upowszechnienie technologii bezprzewodowych, realizowanych w oparciu o technologie 3G/4G sprawiło, że użytkownik zyskał na mobilności oraz przestał być ograniczony dostępnością infrastruktury kablowej.
Dzięki powyższym zmianom użytkownik zyskał łącze o dużej przepustowości, funkcjonujące ciągle, zgodnie z zasadą *Anytime* & *Anywhere* – w dowolnym miejscu i czasie. Pozwoliło to na uruchamianie usług, które były dotychczas niedostępne dla użytkowników „domowych". Były to zarówno usługi wymagające dużej przepustowości, jak i usługi, które wymagały stałego połączenia z siecią Internet. Wdrażanie tych usług było

* Praca naukowa finansowana ze środków na naukę w latach 2007–2009 jako projekt badawczy nr N517 012 32/2108.

procesem ewolucyjnym, który postępował równolegle z rozwojem infrastruktury sieciowej (przewodowej i bezprzewodowej) oraz z rosnącymi możliwościami technicznymi terminali użytkowników. Szczególnym wyzwaniem dla sieci były (i są nadal) te nowo wprowadzane usługi, które wymagają dużych przepustowości i dostarczania różnych rodzajów danych (audio, wideo, tekst, grafika), przesyłanych jednocześnie, w ramach jednej sesji multimedialnej. W niniejszym artykule usługi te będą określane mianem szerokopasmowych usług multimedialnych.

Nowe możliwości zaowocowały pojawieniem się szeregu nowych usług. Część nowo powstających usług szerokopasmowych była przeniesieniem znanych już usług (np. telewizji) do środowiska sieci Internet. Przeniesienie to, choć proste w swej idei, samo w sobie było zagadnieniem nietrywialnym – usługi te były dotychczas realizowane autonomicznie, za pomocą dedykowanych urządzeń końcowych i dedykowanych rozwiązań transmisyjnych. Dedykowane rozwiązania transmisyjne często tworzyły sieć specjalizowaną o złożonej strukturze, zoptymalizowaną pod kątem jednej, konkretnej usługi. Niektóre – jak choćby wspomniana wcześniej telewizja – wymagały dystrybucji przekazu do dużej liczby odbiorców równocześnie. Należy wspomnieć, że problem migracji usług z sieci dedykowanych okazał się niebanalny również z punktu widzenia sieci Internet – taka migracja przekształcała bowiem Internet w środowisko heterogeniczne, wielousługowe i wieloprotokołowe, z różnymi (a czasem nawet sprzecznymi) wymaganiami co do jakości świadczonych usług.

Przeniesienie typowych usług telekomunikacyjnych do Internetu (uogólniając: do środowiska sieci IP) daje jedną, wspólną i spójną infrastrukturę dla wielu usług. Pozwala to ograniczyć koszty, zarówno operatora, jak i użytkownika końcowego. Bywało również, że dzięki nowym metodom transmisji zyskiwano dodatkową funkcjonalność. Tak było choćby w przypadku radia internetowego, które zyskało możliwość przesyłania skojarzonych z dźwiękiem obrazów i (lub) wideo, wykraczając daleko poza ramy tradycyjnej audycji radiowej. Skutkiem opisanej wyżej migracji było pojawienie się w sieci uniwersalnej, jaką jest sieć IP, nowego typu ruchu – ruchu generowanego przez szerokopasmowe usługi multimedialne. W pewnych okolicznościach ruch ten może podlegać niekorzystnemu oddziaływaniu ze strony ruchu pochodzącego od innych, już istniejących lub nowo wprowadzanych, aplikacji. W efekcie można zaobserwować albo niską jakość usług nowo wprowadzanych, albo degradację usług już istniejących, albo jedno i drugie. Warto przy tym zauważyć, że wiele problemów, które nie występowały w dedykowanej sieci homogenicznej, w sieci heterogenicznej pojawia się dopiero po uruchomieniu usługi na dużą skalę.

Niniejszy artykuł ma na celu przedstawienie problemów, jakie występują przy wdrażaniu nowych rozwiązań, oraz wskazanie tych kierunków poszukiwań rozwiązań owych problemów, które wydają się najbardziej obiecujące w świetle badań przeprowadzonych przez Autorów. Rozdział 2 artykułu zawiera analizę szerokopasmowych usług multimedialnych. W rozdziale 3 została przedstawiona analiza heterogenicznych sieci IP. Rozdział 4 omawia problemy, jakie występują przy interakcjach ruchu generowanego przez różne aplikacje. Rozdział 5 poświęcony został problemom związanym z przeciwdziałaniem przeciążeniom i metodom tzw. sprzyjania TCP. Rozdział 6 stanowi podsumowanie niniejszego artykułu.

2. Analiza szerokopasmowych usług multimedialnych

Jak wspomniano w poprzednim rozdziale, zmiany technologiczne obserwowane na przestrzeni ostatnich lat (zwiększenie przepustowości łączy dostępowych, wzrost mocy obliczeniowych, wzrost wydajności urządzeń mobilnych) umożliwiły wprowadzanie do sieci IP szerokopasmowych usług multimedialnych. Jedną z takich usług jest usługa telewizyjna. Oferowana jest ona zarówno jako usługa alternatywna dla istniejącej telewizji (analogowej oraz cyfrowej), jak i usługa poszerzająca obszar dostępności telewizji (np. o środowisko terminali mobilnych). Przykładem usługi alternatywnej jest usługa IPTV (*Internet Protocol Television*), którą ogólnie można określić jako przesyłanie sygnału telewizyjnego w sieciach szerokopasmowych, wykorzystujących w warstwie sieciowej protokół IP. Obecnie najczęściej oferowana jest telewizja w rozdzielczości SD (*Standard Definition*), przy czym obserwowana jest migracja do rozdzielczości HD (*High Definition*). Pojawiają się również rozwiązania oferujące telewizję 3D. Obecnie w Polsce udział IPTV w ogólnym rynku usług telewizyjnych szacuje się na 1–2% [15]. Sieci IPTV są sieciami dedykowanymi. Usługa telewizyjna świadczona przez publiczną sieć Internet określana jest mianem telewizji internetowej [3]. Innymi przykładami usług szerokopasmowych realizowanych przez sieć Internet są radio internetowe i różnego typu usługi telekonferencyjne.

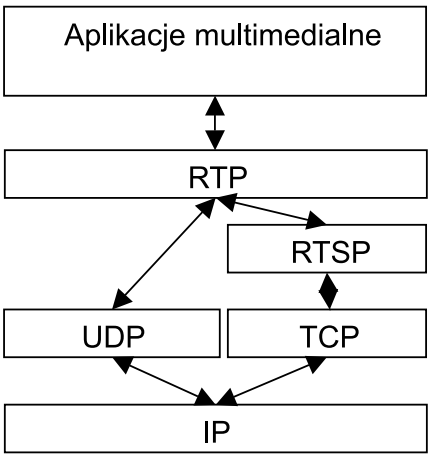
Usługi szerokopasmowe, realizowane w sieci Internet, współpracują z innymi usługami Internetu, takimi jak np. serwis informacyjny WWW i niezawodna transmisja danych masowych. Serwis WWW bardzo często stanowi interfejs użytkownika dla usług szerokopasmowych. Współpracuje on wówczas z protokołami sygnalizacyjnymi usług szerokopasmowych. Natomiast niezawodna transmisja danych innych, niż audio i wideo, jest wykorzystywana m.in. do aktualizacji oprogramowania, przesyłania wiadomości usługi EPG (*Electronic Program Guide*), reklam, itp.

Szerokopasmowe usługi multimedialne wymagają transmisji strumieni audio i (lub) wideo (jednego lub

wielu) oraz, często, przesyłania dodatkowych danych. Przykładowo, radio internetowe typowo wymaga transmisji jednego strumienia audio. Dodatkowo jednak można się spodziewać transmisji złożonego przekazu multimedialnego (zawierającego tekst, grafikę, fotografie, a bywa, że również strumień obrazów). Wynika to z faktu, iż zdecydowana większość stacji radiowych nadaje również reklamy i (lub) dodatkowe informacje związane z nadawaną audycją. Pojawiły się także stacje radiowe nadające oprócz przekazu audio także przekaz wideo – np. popularna „Czwórka", czyli Polskie Radio Program IV, ze swoją ofertą „Radio na Wizji".

W szerokopasmowych usługach multimedialnych transmisja strumieni audio i wideo realizowana jest z wykorzystaniem protokołu transportowego RTP (*Real-time Transport Protocol*) [2] lub, w przypadku serwera strumieniującego firmy *RealNetworks*, protokołu RDT (*Real Data Transport*). Oba te protokoły pozwalają na przesyłanie informacji multimedialnej z zachowaniem warunków czasu rzeczywistego. Zarządzanie transmisją multimedialną realizowane jest często za pomocą protokołu RTSP (*Real-Time Streaming Protocol*) [2].

Protokół RTP współpracuje najczęściej z protokołem transportowym UDP (*User Datagram Protocol*), tworząc stos protokołowy RTP/UDP, gdzie RTP stanowi górną, a UDP dolną podwarstwę warstwy transportowej (rys. 1). Gdy tego typu transmisja nie jest możliwa (np. z powodu stosowania zabezpieczeń typu firewall, których polityka bezpieczeństwa zazwyczaj dyskryminuje protokół UDP) stosowane jest rozwiązanie wykorzystujące protokół RTSP w tzw. trybie *interleaved*. Wówczas RTP pracuje w stosie protokołowym RTP/RTSP/TCP (rys. 1). Przesyłana informacja multimedialna przenoszona jest wewnątrz komunikatów RTSP [4], a zastosowanie protokołu TCP sprawia, że transmisja informacji multimedialnej traci charakter czasu rzeczywistego. Tym niemniej w przypadku sieci nieprzeciążonej, dobrze zwymiarowanej pod kątem przesyłanych multimediów, można osiągnąć efekt zbliżony do czasu rzeczywistego.



Rys. 1. Stos protokołowy wykorzystywany do transmisji multimedialnej

Podstawowym wymaganiem stawianym przez usługi IPTV i telewizji internetowej jest duża przepustowość łączy, odpowiednia do szybkości bitowej przesyłanego strumienia danych. Opóźnienia transmisyjne nie odgrywają znaczącej roli (nawet kilkusekundowe opóźnienia są akceptowalne przez użytkownika). Zmienność opóźnienia powinna być możliwie mała. Jednakże może być ona łatwo kompensowana przez odpowiednio duże bufory. Jest także pożądana niska stopa błędów, aczkolwiek w przypadku telewizji internetowej pojedyncze błędy mogą się zdarzać.

W przypadku radia internetowego należy zachować możliwie małą zmienność opóźnienia (możliwe jest kompensowanie zmienności opóźnienia za pomocą odpowiedniego buforowania). Sygnał audio jest sygnałem o stosunkowo niskiej szybkości bitowej (często poniżej 128 kb/s), stąd duże przepustowości łączy nie są wymagane. Jednakże ze względu na dodatkowe informacje, jakie przesyłają stacje radiowe, pożądane przepustowości są wielokrotnie większe, niż wynikałoby to tylko z szybkości bitowej źródła ruchu audio.

W przeciwieństwie do telewizji i radia, usługi telekonferencyjne mają charakter interaktywny, co sprawia, że jednym z istotniejszych parametrów jakościowych jest opóźnienie transmisyjne. Musi być ono na tyle małe, by pozwoliło na zachowanie interaktywności – czyli nie gorsze niż w klasycznej telefonii (nie więcej niż 150 do 200 ms). Z zagadnieniem interaktywności wiąże się także kolejny parametr jakościowy – zmienność opóźnienia, która musi być bardzo mała. Interaktywny charakter usługi sprawia, że dużych zmienności opóźnienia nie można już kompensować za pomocą buforowania. Przepustowość wymagana dla przekazu dźwięku jest stosunkowo mała (typowo nie przekracza 10 kb/s na jednego uczestnika konferencji). Strumień wideo, w zależności od jakości, z jaką chcemy prowadzić konferencję, waha się od 64 kb/s do 1,5 Mb/s (dla konferencji HD) na uczestnika. W systemach tych pożądana jest także niska stopa błędów.

3. Analiza heterogenicznych sieci IP

Sieć heterogeniczna jest to sieć złożona z podsieci zrealizowanych w różnych technologiach[1]. Encyklopedia

[1] Wprawdzie anglojęzyczny termin „Ethernet (ATM, WDM, etc.) technology" należy słownikowo tłumaczyć jako „technika Ethernet (ATM, WDM, itd.)" (odpowiednikiem terminu „technology" jest polski termin „technika", a nie „technologia"), jednak w wielu publikacjach można napotkać określenie „technologia Ethernet (ATM, WDM, itd.)". Wynika to z faktu, iż technologia (greckie słowo *techne* oznacza „sztukę, rzemiosło", ale również „umiejętność"; logos to „nauka") jest wiedzą o „sztuce" (sposobie, metodzie) wytwarzania lub przetwarzania dóbr (w tym i informacji). Również Autorzy skłaniają się ku opinii, iż termin „technologia" będzie bardziej odpowiedni pojęciowo do opisu ogółu wiadomości na temat sposobu (opisanego konkretnym standardem) przesyłania danych w sieci komputerowej.

PWN definiuje technologię jako „*dziedzinę techniki zajmującą się opracowywaniem i przeprowadzaniem najkorzystniejszych w określonych warunkach procesów wytwarzania lub przetwarzania*". Technologie podsieci składowych opracowywane są pod kątem pracy sieci w określonych warunkach, w których mają one zapewnić efektywną transmisję.

Istniejące technologie sieciowe opracowano do transmisji zarówno na niewielkie odległości, rzędu pojedynczych metrów (sieci PAN, ang. *Personal Area Network*), jak i na znaczne odległości (np. satelitarne sieci VSAT – *Very Small Aperture Terminal*). Charakteryzują się one przepustowościami od 512 b/s (64 B/s), spotykanymi w bezprzewodowych sieciach sensorów [16], do przepustowości przekraczających 10 Tb/s w sieciach DWDM (*Dense Wavelength Division Multiplexing*) [17]. Tak znaczne różnice przepustowości wynikają z optymalizacji technologii pod kątem realizacji określonego typu transmisji w określonych warunkach pracy. Przykładowo, niska przepustowość sieci sensorów wynika z konieczności zapewnienia niskiego poboru mocy (jednego z kluczowych parametrów dla takiej sieci, który wynika z konieczności zapewnienia możliwie długiej pracy na zasilaniu bateryjnym). Innym przykładem może być niewielki zasięg sieci – konieczny dla sieci PAN, który zapewnia w prosty sposób współistnienie na małym obszarze wielu autonomicznych sieci.

Obecnie wśród sieci lokalnych dominują dwa rozwiązania – przewodowa sieć standardu IEEE 802.3 (sieć Ethernet) oraz bezprzewodowa sieć standardu IEEE 802.11 (znana również pod nazwą handlową Wi-Fi). Pierwsza z nich oferuje przepustowości od 10 Mb/s do 10 Gb/s. Druga zapewnia przepustowości od 1 Mb/s do 450 Mb/s. Oprócz nich można spotkać (choć obecnie coraz rzadziej) kilka innych rozwiązań sieci LAN, jak chociażby sieci zgodne ze standardem IEEE 802.5 (*Token Ring*). oferujące przepustowości od 4 Mb/s do 1 Gb/s. Spośród sieci PAN największą popularnością cieszy się obecnie sieć *Bluetooth*.

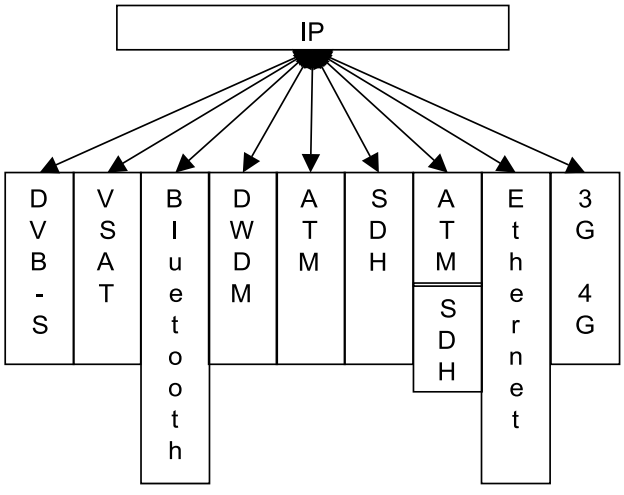
Najpopularniejsze rozwiązania stosowane w sieciach miejskich to ATM i Ethernet 10 Gb/s w przypadku sieci przewodowych, a w przypadku sieci bezprzewodowych – WiMAX (IEEE 802.16). W sieciach rozległych, z kolei, najczęściej spotykane sieci przewodowe to SDH i DWDM, a wśród sieci bezprzewodowych – technologie transmisji danych bazujące na sieciach 3G/4G oraz sieci satelitarne: VSAT, *IP over DVB-S*.

Protokół IP warstwy sieciowej stanowi wspólną platformę Internetu. Łączy on sieci (w sensie warstwy łącza danych i warstwy fizycznej), w jedną strukturę. Protokół IP ma możliwość współpracy z każdą istotną technologią sieciową (rys. 2). Pozwala to na dostęp do Internetu praktycznie w każdym miejscu, w którym można zbudować sieć.

Jednakże tak wielka elastyczność sieci Internet powoduje szereg problemów. Sieć Internet nie jest budowana na bazie sieci homogenicznej. Stanowi połączenie wielu, bardzo różnorodnych sieci, charakteryzujących się różnymi parametrami, zbudowanymi w różnych technologiach sieciowych.

Różnice pomiędzy technologiami zastosowanymi w infrastrukturze Internetu powodują heterogeniczność strukturalną. Tego typu heterogeniczność objawia się m.in. różnicami w rozmiarach MTU (*Maximum Transmission Unit*, maksymalna długość pola danych ramek przenoszących datagramy IP), różnymi wartościami opóźnienia i zmienności opóźnienia, różnymi przepustowościami czy różnicami w sposobie nawiązywania połączenia. Niektóre z tych różnic są niwelowane już w warstwie sieciowej, dzięki mechanizmom protokołu IP (np. mechanizm fragmentacji, realizujący dopasowanie datagramu IP do MTU ścieżki transmisji datagramu). Niemniej jednak podstawowe różnice pomiędzy sieciami pozostają – i, niejednokrotnie bardzo silnie, wpływają na transmisję.

Omawiając heterogeniczność strukturalną nie sposób nie wspomnieć o algorytmach dostępu do medium transmisyjnego. Dostęp rywalizacyjny, spotykany w popularnych sieciach lokalnych, w mocno obciążonej sieci może powodować znaczne wahania przepustowości oraz opóźnienia transmisyjnego, obserwowane przez pojedynczego użytkownika. Z punktu widzenia końcowego odbiorcy przepustowość nie jest stała, a technologia określa tylko jej wartość maksymalną. W przypadku dostępu kontrolowanego (stosowanego np. w sieci *Token Ring* – dostęp kontrolowany z przydziałem uprawnienia) użytkownik uzyskuje mniejsze fluktuacje czasu dostępu do łącza i względnie stałą przepustowość. Dzieje się to jednak kosztem gorszego wykorzystania łącza w przypadku sieci słabo obciążonej oraz większych opóźnień transmisyjnych.



Rys. 2. Współpraca protokołu IP z wybranymi technologiami

W efekcie, w sieciach heterogenicznych, zbudowanych przy wykorzystaniu pewnych technologi, mogą występować znaczne wahania przepustowości związane z mechanizmem dostępu do medium. Wpływa to w sposób istotny na efektywną przepustowość, jaką uzyska użytkownik końcowy. Podobne problemy mogą występować w przypadku opóźnień transmisyjnych. Jest to zjawisko niekorzystne dla usług realizujących transmisję w czasie rzeczywistym. Warto zauważyć, że wybrane technologie częściowo uwzględniają tę problematykę (przykładowo, sieć standardu 802.11 stosuje dwie funkcje koordynacji dostępu do medium - rozproszoną, opartą o dostęp rywalizacyjny, oraz punktową, z dostępem kontrolowanym).

Technologie wpływają na transmisję w sposób quasi-statyczny. Do czynników quasi-statycznych dochodzą czynniki dynamiczne (są to przede wszystkim czynniki związane z dynamiką ruchu w sieci Internet). Wpływ czynników dynamicznych objawia się znacznymi zmianami obciążeń poszczególnych łączy, zależnie m.in. od charakteru ruchu generowanego przez użytkowników sieci. Inny charakter będzie miał ruch generowany podczas transmisji danych masowych, inny podczas sesji WWW, a jeszcze inny będzie generowany przez usługi szerokopasmowe. Natomiast to, jak sieć będzie reagowała na czynniki dynamiczne, będzie zależało od zastosowanych rozwiązań architektonicznych danej sieci (m.in. mechanizmów równoważenia obciążeń, kontroli ruchu przychodzącego, kształtowania ruchu itp.).

Część usług szerokopasmowych wymaga dystrybucji danych do wielu odbiorców. Są to głównie te usługi, które wywodzą się z systemów rozsiewczych (jak np. radio i telewizja), ale należą do nich również np. usługi wykorzystywane do pracy grupowej. Wymaganie to może być spełnione m.in. dzięki zastosowaniu technologii wspierających transmisję multikastową. Transmisja tego typu pozwala na efektywne wykorzystanie dostępnych zasobów sieciowych [1, 2].

W transmisji multikastowej dane są przesyłane od nadawcy do wielu odbiorców równocześnie. Tworzy się w ten sposób nie pojedyncza ścieżka (jak w transmisji unikastowej), ale drzewo dystrybucji (którego korzeniem jest nadajnik, a liśćmi odbiorniki), którego gałęzie w sieci heterogenicznej mogą przechodzić przez fragmenty sieci o różnych właściwościach.

Heterogeniczność sieci powoduje, iż drzewa dystrybucji multikastowej mogą również być heterogeniczne, co utrudnia dystrybucję. Podobnie jak ścieżka transmisji unikastowej, tak i drzewo dystrybucji multikastowej pozwala na transmisję danych z szybkością, na jaką pozwala fragment o najniższej przepustowości. Przepustowość obserwowana dla danego odbiornika jest ograniczona minimalną przepustowością na ścieżce od nadajnika do odbiornika, prowadzonej zgodnie z drzewem dystrybucji.

Usługi szerokopasmowe wymagające transmisji multikastowej muszą uwzględniać heterogeniczność drzewa dystrybucji. Nie można wówczas wysyłać danych do wszystkich odbiorców z szybkością bitową, na jaką pozwala fragment drzewa dystrybucji o najwyższej przepustowości. Trzeba dopasować się do tych odbiorców, w przypadku których droga dystrybucji informacji przebiega przez fragmenty drzewa o najniższej przepustowości. Braki w dopasowaniu objawiają się najczęściej w postaci przeciążeń w tych węzłach, w których szybkość bitowa źródła ruchu przekracza dostępną przepustowość łącza. Szerzej zagadnienie to będzie przedstawione w rozdziale 5.

Heterogeniczność sieci IP jest widoczna nie tylko w warstwach znajdujących się poniżej IP (łącza danych i fizycznej) realizowanych za pomocą różnych technologii sieciowych. Występuje ona także w warstwach ulokowanych powyżej IP (głównie warstwy transportowej i warstwy aplikacji) i związana jest ona z zastosowaniem różnych protokołów tych warstw. Nie jest to jednak heterogeniczność strukturalna, charakterystyczna dla niższych warstw sieci, lecz heterogeniczność wynikająca ze współistnienia w tej samej infrastrukturze (często również heterogenicznej) różnych protokołów i aplikacji. Heterogeniczność mająca swe źródło w wieloprotokołowości i wielousługowości sieci. Ze względu na złożoność tej problematyki, zostaną jej poświęcone dwa kolejne rozdziały.

4. Analiza wzajemnego oddziaływania ruchu generowanego przez nowe i klasyczne usługi Internetowe

Na charakter ruchu generowanego przez usługi sieciowe ma wpływ wiele czynników. Dla ruchu generowanego w trakcie pojedynczej transmisji (podczas przesyłania np. pojedynczego pliku czy pojedynczego strumienia audio) ważny jest m.in. rodzaj przesyłanych danych (skompresowane wideo, pliki binarne, skompresowana grafika, itp.), oraz to, w jakich warunkach (z zachowaniem warunków czasu rzeczywistego, bez zachowania warunków czasu rzeczywistego) i poprzez jaki stos protokołowy (TCP/IP, UDP/IP, RTP/UDP/IP, RTP/RTSP/TCP/IP, itp.) transmisja będzie realizowana.

Usługa sieciowa może wymagać transmisji jednego lub wielu strumieni danych. Usługi szerokopasmowe najczęściej wymagają przesyłania wielu strumieni danych. Innymi słowy, pojedyncza usługa szerokopasmowa może generować ruch będący złożeniem kilku pojedynczych transmisji.

Usługi mogą generować ruch o jednorodnym lub niejednorodnym charakterze. W pierwszym przypadku

może to być jeden lub wiele strumieni przenoszących dane tego samego typu i w identyczny sposób, np. dwa strumienie audio przesyłane za pomocą protokołu RTP. W drugim przypadku może być przesyłanych wiele rodzajów danych (np. audio, wideo, tekst) przez te same lub inne protokoły, ewentualnie ten sam rodzaj danych, ale przez znacząco różniące się protokoły lub z wykorzystaniem mechanizmów znacząco zmieniających charakterystykę przynajmniej jednego ruchu. Przykładem może tu służyć aplikacja multimedialna realizująca transmisję jednego lub wielu strumieni mediów oraz danych, które mogą być związane z zewnętrznym zarządzaniem usługą lub też stanowić integralną część usługi.

W przypadku „klasycznych" usług radia i telewizji, funkcjonujących w dedykowanej sieci (rozsiewczej sieci bezprzewodowej, naziemnej lub satelitarnej), nie występują istotne interakcje pomiędzy poszczególnymi usługami (kanałami radiowymi czy telewizyjnymi). Mimo iż w ramach usługi mogą być przesyłane także dodatkowe dane (np. w transmisji radiowej komunikaty RDS – *Radio Data System*, a w telewizyjnej telegazeta), ta dodatkowa transmisja nie ma istotnego wpływu na podstawowy przekaz. Jest ona wprawdzie przesyłana tym samym kanałem, ale poza podstawową transmisją (RDS – powyżej częstotliwości przesyłanego sygnału akustycznego, telegazeta – w niewyświetlanych liniach obrazu nadawanych w czasie wygaszania powrotów plamki).

Inaczej jest w przypadku transmisji w sieci teleinformatycznej. Jeżeli dowolna usługa, w tym i szerokopasmowa, generuje nie jeden, lecz kilka strumieni, to strumienie te będą na siebie oddziaływać, jeśli nie zapewnimy im dodatkowej separacji. To wzajemne oddziaływanie na siebie strumieni występuje nie tylko w relacji usługa-środowisko sieciowe, ale i w ramach jednej usługi. Jeżeli zatem nawet skorzystamy z sieci dedykowanej, gdzie dla każdej usługi będzie wydzielony kanał transmisyjny (fizyczny lub wirtualny), to i tutaj mogą wystąpić interakcje (wewnętrzne, w ramach jednej usługi). Jeszcze gorsza sytuacja wystąpi w sieci wielousługowej, np. w sieci Internet. Tutaj możliwe są nie tylko interakcje w ramach danej usługi, ale również interakcje pomiędzy takimi samymi usługami, jak i interakcje pomiędzy różnymi usługami.

Generalnie, ruch generowany w sieci Internet przez usługi można podzielić na ruch elastyczny (*elastic traffic*) i ruch nieelastyczny (ang. *inelastic traffic*). Pierwszy jest charakterystyczny dla transmisji danych masowych, drugi dla transmisji informacji multimedialnej w czasie rzeczywistym. W szczególnym przypadku, aplikacje multimedialne mogą również generować ruch elastyczny. Jest to spotykane np. w jednym z rozwiązań stosowanych dla potrzeb transmisji telewizyjnych w Internecie – transmisja RTSP w trybie *interleaved* [4].

Od dawna obserwowane jest zjawisko negatywnego oddziaływania ruchu nieelastycznego, generowanego przez aplikacje multimedialne, na ruch elastyczny generowany przez aplikacje stosujące do transmisji protokół TCP warstwy transportowej. Zjawisko to, określane mianem braku sprzyjania TCP (*TCP – infriendliness*), jest związane z nadmiernym spadkiem przepływności protokołu TCP w sytuacji, gdy transmisja TCP rywalizuje o zasoby sieciowe z protokołem (protokołami), które nie implementują mechanizmu przeciwdziałania przeciążeniom podobnego w działaniu do zaimplementowanego w TCP. Często towarzyszy temu brak, typowego dla konkurujących ze sobą połączeń TCP, sprawiedliwego (w sensie: równego) podziału zasobów sieciowych pomiędzy rywalizujące połączenia. W celu wyeliminowania (a przynajmniej znacznego ograniczenia) tego zjawiska, niekorzystnego z punktu widzenia pracy sieci, została opracowana koncepcja systemu (protokołu, architektury) sprzyjającego TCP (*TCP-friendly*) [1].

W sieciach dobrze zwymiarowanych pod kątem transmisji informacji multimedialnej w czasie rzeczywistym, transmisja multimedialna realizowana w stosie protokołowym RTP/UDP/IP nie powoduje braku sprzyjania TCP [9]. W takiej sieci straty podczas transmisji informacji multimedialnej są na dopuszczalnym poziomie, a przepływność protokołu RTP jest akceptowalna. Dodatkowo, zastosowanie w transmisji multimedialnej mechanizmu ograniczającego rozmiar zgęstki pakietów i zmieniającej jej charakter [8] poprawia współistnienie obydwu rodzajów ruchu. Mechanizm zaproponowany w [8] znacznie poprawia współistnienie ruchu elastycznego i nieelastycznego, zapewniając zachowanie warunków czasu rzeczywistego podczas transmisji informacji multimedialnej.

Inną często wskazywaną możliwością poprawy współistnienia ruchu elastycznego i nieelastycznego jest użycie do transmisji strumienia multimedialnego któregoś z protokołów sprzyjających TCP, np. protokołu TFRC (*TCP Friendly Rate Control*). Zastosowanie protokołu sprzyjającego TCP jest o tyle korzystne, że pozwala uniknąć zjawiska braku sprzyjania TCP [6]. Jednakże w warunkach, w których zachowanie sprzyjania TCP pozostaje w konflikcie z warunkami czasu rzeczywistego, protokół TFRC (wskazywany często w literaturze jako protokół sprzyjający TCP odpowiedni dla transmisji multimedialnej), nie był w stanie zapewnić warunków transmisji czasu rzeczywistego i tym samym uniemożliwiał poprawne odtwarzanie przekazu multimedialnego [6].

Na współistnienie ruchu elastycznego i nieelastycznego ma również wpływ typ kolejek stosowanych w sieci. Podstawową metodą zarządzania kolejką jest *Tail Drop* – czyli bufor FIFO z procedurą obsługi, która w przypadku przepełnienia bufora odrzuca nadchodzące pakiety. Innym rozwiązaniem jest kolejka typu RED.

W kolejce RED decyzja o odrzuceniu pakietu jest podejmowana na podstawie przeciętnego zapełnienia bufora. W kolejce tej pojedyncze pakiety są odrzucane najczęściej jeszcze przed przepełnieniem się kolejki. Dla strumieni poddających się przeciwdziałaniu przeciążeniom, takich jak TCP, strata pakietu jest informacją o wystąpieniu w sieci przeciążenia. Mechanizm zapobiegania i przeciwdziałania przeciążeniom zastosowany w protokole TCP powoduje wtedy spowolnienie tempa wysyłania danych, tak by umożliwić rozładowanie natłoku w sieci.
Przy dużym przeciążeniu, ruch nieelastyczny może zmonopolizować całą przestrzeń bufora i całkiem zdegradować transmisje TCP. W pewnych warunkach, zastosowanie aktywnego zarządzania kolejką typu RED może zmniejszyć negatywny wpływ, jaki ruch nieelastyczny ma na strumienie elastyczne [7]. Dodatkowo kolejka typu RED pozwala na uniknięcie synchronizacji wielu strumieni TCP i uniknięcie efektu *Lock Out*[2].

5. Analiza metod zapobiegania i przeciwdziałania przeciążeniom w heterogenicznych sieciach IP

W heterogenicznych sieciach IP przeciążenia są zjawiskiem typowym, chociaż niepożądanym. Dlatego bardzo ważne są mechanizmy zapobiegające powstawaniu nowych przeciążeń oraz mechanizmy rozładowywania przeciążeń już istniejących.
Jak wspomniano w rozdziale 3., heterogeniczność sieci IP rozumiemy dwojako. Z jednej strony, jako połączenie sieci zbudowanych w oparciu o różne technologie w jedną całość, jedną wspólną sieć IP. Z drugiej strony, jako sieć niejednorodną (wieloprotokołową) w wyższych warstwach (od 4 do 7) modelu OSI.
Stosowane w sieci Internet technologie często w sposób znaczący różnią się pod względem oferowanej przepustowości. Takie różnice najczęściej prowadzą do powstawania przeciążeń (często permanentnych) na granicy pomiędzy sieciami o różnych przepustowościach, na kierunku od sieci szybszej (o wyższej przepustowości) do wolniejszej (o niższej przepustowości). Najwolniejszy fragment sieci na ścieżce dystrybucji unikastowej (lub drzewie dystrybucji multikastowej) stanowi wąskie gardło systemu. Granica wąskiego gardła, czyli styk pomiędzy najwolniejszym a szybszym fragmentem sieci, jest miejscem, gdzie występuje gros zjawisk związanych z zaburzeniami transmisji – przeciążenia, wynikające z nich zwiększone opóźnienia i niekontrolowane straty pakietów. W sieciach dobrze zwymiarowanych pod kątem przenoszonego ruchu nie występują permanentne przeciążenia związane z różnicą technologii.
Wielousługowość sieci oraz wynikająca z niej wieloprotokołowość wyższych warstw sieci prowadzi do przeciążeń (lub zjawisk identyfikowanych przez mechanizmy sieciowe jako przeciążenia), które w przypadku usług szerokopasmowych wynikają głównie ze współistnienia ruchu elastycznego i nieelastycznego w łączach współdzielonych. Najogólniej rzecz biorąc, problemy, z jakimi wówczas boryka się sieć, można podzielić na związane ze strukturą ruchu i charakterem transmisji oraz związane z interakcjami pomiędzy odmiennymi mechanizmami stosowanymi w protokołach. Rozpatrując zagadnienie przeciwdziałania i zapobiegania przeciążeniom są to, odpowiednio, problemy związane z powstawaniem chwilowych przeciążeń oraz związane z różną reakcją protokołów na przeciążenia. Powstawanie chwilowych przeciążeń wynikających ze struktury ruchu i charakteru transmisji ogranicza się stosując różnego typu metody kształtowania ruchu (w tym wygładzanie ruchu). Przykładowo, w sieciach dobrze zwymiarowanych pod kątem transmisji informacji multimedialnej w czasie rzeczywistym, transmisja multimedialna realizowana w stosie protokołowym RTP/UDP/IP z zastosowaniem mechanizmu zapewniającego tolerancję dla TCP (ograniczającego rozmiar zgęstki pakietów) poprawia współistnienie ruchu elastycznego i nieelastycznego [8]. Badania prowadzone pod kątem zapobiegania przeciążeniom wyjaśniły charakter tego zjawiska [14]. Wykazano m.in., że mechanizm wygładza charakterystykę czasową zajętości bufora, likwiduje chwilowe przeciążenia wynikające ze struktury ruchu multimedialnego, co jest obserwowane jako znaczne zmniejszenie wariancji zajętości bufora. Przeciążenia wynikające ze struktury ruchu ulegają naturalnemu rozładowaniu bez udziału mechanizmów protokołowych przeciwdziałających przeciążeniom. Jak wykazały badania, są one jednak mylnie interpretowane przez te mechanizmy (które traktują je jako przeciążenia wymagające reakcji), co prowadzi do niepotrzebnego ograniczania przepływności protokołów implementujących dany mechanizm (takich jak np. protokół TCP). Zastosowanie mechanizmu tolerancji dla TCP, zapobiegającego powstawaniu przeciążeń wynikających ze struktury ruchu, pozwoli nie tylko uniknąć strat pakietów wynikających z przeciążeń chwilowych, ale również nie dopuszcza do niepotrzebnego uruchamiania mechanizmów protokołowych i architektur przeciwdziałania przeciążeniom i, w efekcie, do niepotrzebnego ograniczania przepływności TCP [14].

[2] Zjawisko to charakteryzuje się tym, że pakiety od jednego bądź kilku nadawców zawsze docierają do urządzenia w momencie, kiedy bufor kolejki jest już zapełniony, co uniemożliwia transmisję.

Problemy związane z interakcjami pomiędzy odmiennymi mechanizmami stosowanymi w protokołach dotyczą różnej reakcji protokołów na przeciążenia. Jednym z najbardziej znanych problemów tego typu jest brak sprzyjania TCP. Skutkiem tego jest obecna tendencja w budowie protokołów transportowych i architektur zapobiegania przeciążeniom, by były one „sprzyjające TCP" [1], a zatem, by w sposób rozsądnie sprawiedliwy (*reasonably fair*) chroniły ruch TCP przy jednoczesnym zapewnieniu QoS dla transmitowanego ruchu.

Spośród protokołów sprzyjających TCP, jako najbardziej odpowiedni do zapobiegania przeciążeniom podczas transmisji informacji multimedialnej w czasie rzeczywistym jest wskazywany protokół TFRC. Wyniki badań przeprowadzonych pod kątem przeciwdziałania przeciążeniom wykazały, że protokół TFRC skutecznie przeciwdziałał przeciążeniom w całym zakresie analizowanych warunków pracy. Należy jednak zwrócić uwagę na wyniki badań tego protokołu pod kątem współistnienia ruchu elastycznego i nieelastycznego [6]. Wykazały one, że w pewnych sytuacjach protokół TFRC nie był w stanie zapewnić warunków przesyłania w czasie rzeczywistym i tym samym uniemożliwiał poprawne odtwarzanie przekazu multimedialnego.

Możliwym rozwiązaniem tego problemu jest, zaproponowana w [5], modyfikacja protokołu TFRC. Zastosowane w TFRC nieliniowe równanie (modelujące zachowanie protokołu TCP) zastąpiono równaniem liniowym (określającym przepływność protokołu RTP w funkcji błędów). Pozwoliło to na uzyskanie „łagodnego" przeciwdziałania przeciążeniom, co umożliwiło pogodzenie konieczności ograniczenia szybkości bitowej z koniecznością zachowania warunków czasu rzeczywistego. Kombinacja „łagodnego" przeciwdziałania przeciążeniom oraz tolerancji dla TCP pozwoliła z kolei na pogodzenie wymagań ruchu elastycznego i nieelastycznego. W efekcie, ruch multimedialny jest przesyłany z zachowaniem warunków czasu rzeczywistego, a protokół TCP zajmuje przepustowość niewykorzystaną przez ruch multimedialny.

Ostatnim z omawianych zagadnień jest konieczność zapewnienia specjalizowanych mechanizmów do przeciwdziałania przeciążeniom pojawiającym się w heterogenicznym drzewie dystrybucji multikastowej. Przeciążenia takie najczęściej nie mogą być rozwiązywane „klasycznymi' metodami przeciwdziałania przeciążeniom, zorientowanymi na transmisję unikastową, gdyż prowadziłoby to do niepotrzebnego ograniczania przepustowości w nieprzeciążonych gałęziach drzewa. Dlatego w przypadku multikastowej transmisji informacji multimedialnej w czasie rzeczywistym, zapobieganie i przeciwdziałanie przeciążeniom wymaga najczęściej stosowania specjalizowanych architektur, radzących sobie z problemem heterogeniczności technologii.

W sieciach IP, w których występuje zarówno heterogeniczność technologii sieciowej, jak i heterogeniczność ruchu, zjawiska wynikające z obydwu tych heterogeniczności nakładają się na siebie. Oznacza to, iż proponowane rozwiązania powinny dodatkowo zapewniać sprzyjanie TCP lub przynajmniej tolerancję dla TCP.

Zastosowanie w tym przypadku protokołu sprzyjającego TCP, jak TFMCC (*TCP-Friendly Multicast Congestion Control* – multikastowy wariant protokołu TFRC) czy PGMCC, niepotrzebnie obniża przepływność we wszystkich gałęziach drzewa dystrybucji multikastowej, dopasowując ją do przypadku najgorszego. Ogranicza to stosowalność multikastowych protokołów sprzyjających TCP.

Najkorzystniejsze są rozwiązania bazujące na multikastowej transmisji warstwowej lub replikacji strumieni. Jednym z protokołów implementujących transmisję warstwową jest WEBRC (*Wave and Equation Based Rate Control*) [1]. Innym z należących do tej klasy rozwiązań jest system dystrybucji informacji multimedialnej w czasie rzeczywistym, zbudowany w oparciu o architekturę multikastowej transmisji warstwowej współpracującej z sygnalizacją ECN (*Explicit Congestion Notification*) [12]. Opisany w [12] mechanizm zapewnia sprzyjanie TCP, zapewniając równocześnie zachowanie właściwych charakterystyk transmisji informacji multimedialnej w czasie rzeczywistym. Mechanizm ten może być wykorzystywany także do transmisji danych masowych przy wykorzystaniu protokołów ALC (*Asynchronous Layered Coding*) w warstwie transportowej i FLUTE (*File Delivery over Unidirectional Transport*) w warstwie aplikacji [1].

6. Podsumowanie

W artykule omówiono problemy występujące przy wdrażaniu nowych usług do heterogenicznych sieci IP oraz możliwości ich rozwiązania. Pod uwagę wzięto zarówno heterogeniczność w niższych warstwach modelu OSI/ISO (heterogeniczność technologii), jak i heterogeniczność w wyższych warstwach (heterogeniczność protokołów transportowych i usług). Problemy te przejawiają się w dwóch aspektach: współistnienie ruchu elastycznego i nieelastycznego w łączu współdzielonym oraz zapobieganie i przeciwdziałanie przeciążeniom.

Literatura

[1] Chodorek A., Chodorek R.R., Pach A.R.: *Dystrybucja danych w sieci Internet*. WKŁ, Warszawa 2007

[2] Chodorek R.R., Pach A.R.: *Transmisja multikastowa w sieciach IP*. Wydawnictwo Fundacji Postępu Telekomunikacji, Kraków 2003

[3] Krempa A.: *Analiza transmisji telewizji internetowej w Polsce*. Raport Katedry Telekomunikacji 1/2007, AGH, Kraków 2007

[4] Chodorek A., Chodorek R.R.: *Transmisja informacji multimedialnej przez TCP – schemat transmisji RTSP INTERLEAVED*. Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne, R. 81, nr 4, s. 229–232

[5] Chodorek A., R.R. Chodorek R.R.: *Streaming video over TFRC with linear throughput equation*. Materiały konferencji PWT'2007, Poznań, 6–7 grudnia 2007

[6] Chodorek A., Chodorek R.R.: *Applicability of TCP-friendly protocols for real-time multimedia transmission*. Materiały konferencji PWT'2007, Poznań, 6–7 grudnia 2007

[7] Krempa A.: *Analysis of RED algorithm with responsive and non-responsive flows*. Materiały konferencji PWT'2007, Poznań, 6–7 grudnia 2007

[8] Chodorek A., Chodorek R.R.: *An analysis of TCP-tolerant real-time multimedia distribution*. Materiały konferencji: HET-NETs 2008, s. 1–8

[9] Chodorek A., Chodorek R.R., Krempa A.: *An analysis of elastic and inelastic traffic in shared link*. Materiały konferencji: Conference on Human system interaction, HSI 2008, s. 873–878

[10] Krempa A.: *Dual queue management as a method of improvement of coexistence of different types of traffic*. Materiały konferencji: PWT 2008, Poznań, 11 grudnia 2008, s. 1–4

[11] Krempa A.: *Separation of elastic and inelastic traffic and its impact on protocols performance*. Materiały konferencji: PWT 2008, Poznań, 11 grudnia 2008, s. 1–4

[12] Chodorek A., Chodorek R.R.: *ECN-capable TCP-friendly layered multicast multimedia delivery*. Materiały konferencji: Eleventh international conference on Computer modeling and simulation: Cambridge, UK, UKSIM 2009 March 25–27, 2009, s: 553–558

[13] Krempa A.: *Analysis of different approaches in separation of elastic and inelastic traffic*. Materiały konferencji: Eleventh international conference on Computer modeling and simulation: Cambridge, UK, UKSIM 2009 March 25–27, 2009, s: 508–513

[14] Chodorek A., Chodorek R.R.: *An Analysis of TCP-Tolerant Real-Time Multimedia Distribution in Heterogeneous Networks*. Rozdział w książce: pod red. D.D. Kouvatsosa: „Traffic Engineering, Performance Evaluation Studies and Tools for Heterogeneous Networks", River Publishers, Gistrup, Denmark 2009, s. 315–336 (ISBN: 978-87-92329-16-5)

[15] *Raport z badania rynku usług dostępu do płatnej telewizji*. Urząd Ochrony Konkurencji i Konsumentów, Departament Analiz Rynku, Sierpień 2011, http://www.uokik.gov.pl/download.php?plik=10701

[16] Callaway E.H.: *Wireless Sensor Networks: Architectures and Protocols*. CRC Press, Inc., Boca Raton, FL, 2003

[17] Su Yang, Xu Zhanqi, Zhao Ruiqin, Liu Zengji: *Study on Strategy of Dynamic Joint Routing and Resource Allocation in Layered Optical Transport Networks*. Journal of Electronics (China), No. 2, 2008, s. 166–173