1.4 Podpora multimediálních aplikací v Internetu

1.4.1 Požadavky multimediálních aplikací na kvalitu služby (QoS)

Uspokojivá funkce současných multimediálních aplikací realizujících přenosy dat, hlasu i videa včetně konferencí v reálném čase je závislá na kvalitě přenosové služby poskytované počítačovou sítí. Donedávna nebyla kvalita služby v Internetu příliš řešena a směrování paketů se dělo výhradně metodou ”best effort“, kdy se síť sice snažila každý paket doručit k cíli, avšak v případě zahlcení nijak nerozlišovala mezi pakety důležitějšími a méně důležitými. V poslední době je snaha zabezpečit mechanismy, které dovolí provoz v síti priorizovat, nebo dokonce rezervovat přenosové pásmo pro určité konkrétní datové toky. Je důležité si uvědomit, že kvalita služby je ovlivněna všemi komponentami sítě, a to zejména

- stanicemi (pracovními stanicemi i servery)
- síťovými prvky (směrovači, přepínači)
- linkami (spojovací linky mezi směrovači, technologie LAN segmentů)

Podporu mechanismů zajišťujících dodržení parametrů kvality služby tak můžeme najít na všech vrstvách modelu OSI-RM.

1.4.1.1 Parametry QoS v paketových sítích

Kvalita služby (nebo-li QoS z anglického „Quality of Service“) chápeme jako schopnost sítě podporovat aplikaci bez omezení funkce nebo výkonu aplikace. Formálně definuje norma ITU-T E.800 kvalitu služby jako "souhrnný výsledek výkonnosti služby, který určuje stupeň spokojenosti uživatele služby". Z technického hlediska je kvalita služby měřitelná zejména hodnotami následujících parametrů:

- šířka pásma (bandwidth)
- zpoždění (delay)
- rozptyl zpoždění (jitter)
- ztrátovost paketů (packet loss)

Šířka pásma je limitována nejméně propustnou linkou mezi komunikujícími zařízeními, ale také mírou zatížení síťových prvků na cestě. Zpoždění v síti je způsobeno dvěmi příčinami. Pevná složka zpoždění je dána jednak dobou šíření signálu médiem a jednak tzv. serializačním zpožděním, které vznikne tak, že při průchodu paketu síťovým prvkem (směrovačem či přepínačem) je třeba vždy paket nejprve celý přečíst do paměti, rozhodnout o výstupním rozhraní a pak opět z paměti serializovat bit po bitu na výstupní linku. Zpoždění v každém síťovém prvku tak bude dáno podílem délky paketu a přenosové rychlosti[1]. Mimo pevné složky má zpoždění i složku proměnnou, která je způsobena čekáním paketu ve frontách jednotlivých síťových prvků na cestě. Právě různý způsob zacházení s pakety v jednotlivých síťových

[1] U některých přepínačů sice není třeba rámec načítat celý, čekání na načtení alespoň části hlavičky obsahující cílovou adresu se však před zahájením předávání rámce na výstupní port nevyhneme.

prvcích co do obsluhy front je jednou ze základních příčin rozptylu zpoždění, jehož vysoké hodnoty jsou nepříjemné zejména pro multimediální aplikace v reálném čase. Sjednocení způsobu zacházení s pakety určité třídy provozu ve všech síťových prvcích je jedním z postupů, jak vylepšit kvalitu služby poskytovanou paketovou sítí.

Příčinou ztrátovosti paketů jsou jednak chyby přenosu na fyzické vrstvě, může však jít i o zahazování paketů síťovými prvky. Síťové prvku typicky pakety zahazují buďto při přetížení procesoru přerušeními od vstupního provozu (tzv. input drops) nebo při přeplnění výstupních front (tzv. output drops). Ztrátovost paketů by měla být co nejmenší, multimediální aplikace jako přenos hlasu a videa jsou však k určitým ztrátám paketů tolerantní. Krátké výpadky totiž nemusí uživatel zaznamenat, nebo lze malé množství vypadlých vzorků přenášeného signálu dokonce částečně zrekonstruovat.

1.4.1.1.1 Omezování a tvarování provozu

Před tím, než jsou pakety vůbec vpuštěny do sítě a předloženy k rozhodnutí, s jakou preferencí mají být mají být síťovými prvky přeposílány, jsou na hranici sítě aplikovány mechanismy omezování provozu (traffic policing) s ohledem na to, jakou přenosovou kapacitu má uživatel s provozovatelem své sítě nebo poskytovatele Internetu smluvenu. Zde si je však třeba uvědomit, že charakter provozu u dnešních klient-server aplikací je velmi shlukový a často asymetrický. Proto se zpravidla nedefinuje jen maximální bitová rychlost, ale spíše jen množství bajtů, které je uživatel průměrně oprávněn přenést za určitý časový interval a dále míra shlukovitosti, jakou provoz může vykazovat (tj. do jaké míry je možné množství bajtů, které nebyly přeneseny v předchozích intervalech přenést navíc v intervalu současném). Traffic policing je zpravidla realizován pomocí algoritmu Token Bucket [5].

Jelikož správcové uživatelských sítí připojených k poskytovateli znají úmluvy se svým poskytovatelem a tedy mohou odhadnout, jaký provoz již traffic policing poskytovatele nebude propouštět, je pro ně výhodné provoz na lince k poskytovateli tvarovat (tzv. traffic shaping). Tvarování provozu spočívá v pozdržení některých (typicky méně prioritních) paketů ve frontách tak, aby byly nepřípustné shluky rozloženy do delšího časového intervalu. Tím se sice zvětší zpoždění některých méně prioritních paketů, avšak zamezí se paušálnímu zahazování všech paketů mechanismem traffic policingu u poskytovatele, jelikož uživatelem správně tvarovaný provoz již poskytovatelem vyžadovaným kritériím vyhoví.
Tvarování provozu je obzvláště užitečné v situacích, kdy fyzická přenosová rychlost rozhraní přesahuje dohodnutou průměrnou přenosovou rychlost. Jelikož podstatou tvarování provozu je dočasné pozdržování vysílaných paketů v bufferech směrovače, je nutné, aby příslušný směrovač disponoval odpovídajícím množstvím paměti RAM.

1.4.1.2 Intserv a Diffserv - modely integrovaných a rozlišovaných služeb

V současných paketových sítích implementujeme dva základní modely pro zajištění kvality služby -- rozlišované služby (differentiated services) a integrované služby (integrated services). Třetím

modelem je původní strategie „best effort“ (absence podpory QoS) historicky dříve používaná v Internetu, kde se síť sice snaží o přenesení paketu, avšak bez jakékoli priorizace při vysílání paketů na zahlcenou výstupné linku i při volbě paketu, který bude zahozen, pokud se přeplňuje fronta příslušného rozhraní.

1.4.1.3 Integrované služby (intserv)

Model integrovaných služeb (integrated services, IntServ) je založen na explicitní rezervaci zdrojů sítě (kapacity linek, paměti ve frontách, podílu času CPU síťových prvků) pro jednotlivé datové toky ještě před zahájením přenosu. Je tedy nutná podpora explicitní rezervace v operačním systému, resp. v aplikačním software koncových stanic. Požadavky na rezervaci přenosového pásma stanice signalizují síťovým prvkům pomocí protokolu RSVP (Resource Reservation Protocol) . Model intserv je přirozený pro spojově orientované sítě (např. ATM), není však příliš vhodný pro sítě paketové. Důvodem je jeho malá škálovatelnost, kdy přes páteřní prvky typicky prochází značné množství (často i krátce trvajících) datových toků a tyto prvky nemají kapacitu udržovat pro všechny procházející toky stav související s rezervací pásma a dostatečně rychle obsluhovat rezervační požadavky pro desítky tisíc neustále vznikajících a zanikajících datových toků.

1.4.1.3.1 Rezervace zdrojů a protokol RSVP

Protokol RSVP (RFC2205) definuje způsob signalizace od příjemce datového toku postupně k aktivním prvkům na cestě ke zdroji určitého datového toku. Požadavek na rezervaci pásma se tedy šíří od příjemce toku proti směru toku a týká se konkrétního datového toku, ať již určeného pro unicast adresu nebo multicastovou skupinu. Rezervace může být unikátní pro jeden konkrétní tok (distinct reservation) nebo sdílená pro skupinu toků (shared reservation).

1.4.1.4 Diffserv

Z důvodu špatné škálovatelnosti modelu Intserv je v dnešní době mnohem rozšířenější model Diffserv. Na rozdíl od Intserv model Diffserv negarantuje absolutní dobu doručení dat, ale řeší pouze upřednostnění některých paketů před jinými (statistická preference). Při dostatečně dimenzovaných linkách a omezení provozu vstupujícího do sítě na úroveň, pro kterou jsou linky dimenzovány, zajišťuje model Diffserv požadovanou kvalitu služby, aniž by trpěl omezeným množství současných toků, jelikož na rozdíl od modelu Intserv neudržuje Diffserv pro jednotlivé datové toky žádnou stavovou informaci.

Principem Diffserv je klasifikace provozu do tříd na hranici sítě (nebo přímo na zdrojové stanici) označováním paketů podle třídy, do které náleží a jejich další zpracování na základě této značky. Značka třídy je připojena do hlavičky paketu (resp. rámce). Tříd bývá obvykle jen několik (zpravidla ne více než deset), takže problém škálovatelnosti zde nevyvstává. Pro každou třídu mají všechny síťové prvky definováno, jak mají s pakety patřící do dané třídy zacházet co do režimu uchovávání ve frontách (tzv. Per-Hop Behavior - PHB). Výhodou tohoto mechanismu je, že pakety jsou klasifikovány do tříd pouze jednou na hranici sítě a vnitřní prvky sítě již dále nemusí paket

zkoumat, pouze aplikují příslušný PHB podle značky obsažené v hlavičce paketu. Také není třeba řešit příslušnost paketů k tokům a pakety lze zpracovávat jednotlivě jen na základě značky.

1.4.1.4.1 Klasifikace provozu

Jak jsme se již zmínili, principem Diffserv je klasifikace a značkování provozu již na vstupu do sítě, jejíž vnitřní prvky již označkování paketů důvěřují a podle značky na ně aplikují příslušné PHB. Klasifikaci lze realizovat buďto přímo v operačním systému stanice nebo na přístupovém prvku sítě (přepínači a následně na směrovači nebo integrovaně na L3 přepínači). Výhodou značkování stanicí je, že se děje přímo u zdroje dat (aplikace), který sám může nejlépe rozhodnout, do jaké třídy data zařadit a jak je označkovat. Na druhou stranu však může docházet k tomu, že uživatelé budou hledat způsoby, jak veškerý svůj provoz označkovávat jako prioritní a tím si vynucovat lepší kvalitu služby i pro své méně důležité aplikace na úkor jiných uživatelů. Proto se v praxi častěji realizuje značkování provozu na přístupovém síťovém prvku. To však představuje pro síťový prvek zvýšenou zátěž, jelikož musí sledovat provoz od uživatele, podle obsahu rozpoznávat použitou aplikaci a na základě toho jednotlivé pakety odpovídajícím způsobem značkovat (tzv- Network-Based Application Recongition). Použitelnost tohoto řešení je ovlivněna množstvím aplikací, které je síťový prvek schopen rozpoznávat a zda lze uživatelsky doplňovat signatury dalších aplikací. Klasifikovat provoz lze obecně podle informací z třetí (někdy i druhé) až sedmé vrstvy modelu OSI RM (např. protokolu, IP adres, TCP/UDP portů, URL, MIME typu, ...) nebo podle rozhraní, kterým provoz do síťového prvku vstupuje.

Při implementaci QoS si je třeba uvědomit, že PHB pro pakety určité třídy provozu je třeba respektovat po celé trase, tedy jak ve směrovačích intranetu nebo sítě WAN , tak v přepínačích koncových LAN segmentů. Jelikož standardní přepínače zpracovávají pouze hlavičku druhé vrstvy OSI RM, musí být pro ně značka určující třídu QoS umístěna do hlavičky 2. vrstvy. Protože původní hlavička rámce Ethernet žádné pole pro identifikaci třídy QoS neobsahovala, byl normou IEEE 802.1p definován nový standard, který před začátek dat dovolí vložit dvoubajtovou pomocnou hlavičku, ve které jsou vyhrazeny 3 bity pro specifikace jedné z 8 tříd Class of Service (CoS). Přítomnost rozšiřující hlavičky je indikována vyhrazenou hodnotou 0x8100 v poli EtherType (obr. 1.3) s tím, že původní hodnota EtherType je zopakována za přídavnou hlavičkou. Všimněme si, že rozšiřující hlavička je současně využívána pro určení čísla VLAN, pokud přenášíme rámec přes trunk linku.

Na třetí vrstvě (tj. v hlavičce IP paketu) je třída provozu identifikována hodnotou v 8-bitovém poli DSCP (Differentiated Service Code Point) . Původně se jednalo o 3- bitové pole Type of Service (ToS) umožňující zvolit jednu z předdefinovaných 8 tříd IP Precedence, později však bylo toto pole při zachování zpětné kompatibility hodnot rozšířeno a nyní lze mimo původních 8 tříd použít ještě 64 různých dalších tříd provozu. Do pole DSCP však není možné vložit libovolnou hodnotu, jelikož je dále strukturováno – rozlišují se zde dvě kategorie služeb - Expedited Forwarding a Assured Forwarding. Expedited Forwarding (RFC2598) slouží jako jakási "virtuální pevná linka", tedy služba konec-konec s malými ztrátami, malým, ale proměnným zpožděním a garantovanou šířkou pásma. V rámci Assured Forwarding (RFC2597) jsou definovány 4 třídy a v z nich 3 precedence pro zahazování paketů daného typu provozu v případě potřeby. Zmíněné strukturování je však spíše pouze konvence, jak třídy provozu na síti rozdělovat - pro pochopení funkčnosti Diffserv není podstatné a proto se jim dále nebudeme zabývat - detaily lze najít např. v [4]

Když paket vstupuje do oblasti sítě podporující QoS, musí být označkován na 2. i na 3. vrstvě. Značka na 2. vrstvě umožní správně priorizovaný průchod rámce přepínanou částí sítě (typicky LAN), značka na 3. vrstvě pak odpovídající průchod směrovanou částí sítě po vybalení paketu z rámce na lokálním přístupovém směrovači na hranici páteře. Na směrovači na hranici cílové lokální sítě pak paket musí být opět zabalen do rámce označkovaného příslušnou prioritou podle 802.1p, aby i v posledním úseku sítě bylo s paketem zacházeno podle požadavků na příslušnou třídu provozu. Přístupový směrovač cílové sítě tak musí podle hodnoty DSCP v IP hlavičce vytvořit hodnotu CoS, kterou vloži do rámce, jenž ponese paket k cílové stanici. Mapování DSCP na CoS není jednoznačné, jelikož značky na 2. a 3. vrstvě mají rozdílný počet bitů. Mapování DSCP na CoS je obvykle na směrovači konfigurováno staticky.

1.4.1.4.2 Mechanismy priorizace provozu

Dokud není linka, na kterou mají být síťovým prvkem směrovány pakety různých tříd provozu zcela vytížena, není třeba mechanismy QoS vůbec aplikovat, protože by pouze přinášely zbytečnou režii[2]. Pakety jsou tedy na linku zasílány podle pravidla FIFO (First In-First Out). Pokud se však začíná plnit výstupní fronta některé linky, nastupují algoritmy, jež mají za cíl určit, který z paketů bude z fronty jako další vybrán a vyslán na linku. Cílem mechanismů zajištění QoS je pro datový tok garantovat alespoň určitou minimální šířku pásma, shora omezené zpoždění a minimální rozptyl zpoždění (jitter).

Technicky se priorizace provozu realizuje zavedením více front pro pakety odcházejících na každou jednotlivou linku a aplikací vhodného režimu obsluhy těchto front. Zpravidla se konfiguruje, které třídy provozu mají být zařazovány do které fronty. Provoz explicitně nevyjmenovaný bývá zařazován do implicitní (default) fronty. Nejčastěji se můžeme setkat s těmito režimy obsluhy front:

- Priority Queuing (PQ) – Fronty mají absolutní priority. Z fronty z větší prioritou se vybírá, dokud není prázdná, až pak může dojít na pakety čekající ve frontě s nižší prioritou. Front může být několik s prioritami od nejvyšší do nejnižší. Nevýhodou PQ je možnost tzv. vyhladovění (starvation) paketů ve frontách s nižšími prioritami, protože dokud tečou jakákoli data s prioritou vyšší, na data s nižší prioritou nikdy nedojde.
- Custom Queuing (CQ) – Pakety se střídavě (round robin) odebírají cyklicky ze všech front. U každé fronty je definováno, kolik bajtů se z ní smí při jednom průchodu nejvýše odebrat. Jedná se v podstatě o proporcionální cyklické přidělování kapacity každé třídě provozu
- Weighted Fair Queuing (WFQ) – Síťový prvek automaticky vyhledává datové toky v příchozím provozu, pakety toků zařazuje do dynamicky vytvářených front a tyto fronty cyklicky obsluhuje. K jednotlivým datovým tokům (určenými protokolem 4. vrstvy, IP adresami a porty transportní vrstvy komunikujících stran) se tak síťový prvek chová férově ve smyslu shodného přidělování částí pásma zahlcené linky. WFQ navíc preferuje kratší pakety, což vede k žádoucí preferenci interaktivního provozu. Slovo „weighted“ (váhovaný) v názvu metody říká, že podíl pásma přídělovaným jednotlivým datovým tokům je navíc ovlivněn hodnotou DSCP v paketech jednotlivých toků.

[2] až na výjimky, jako např. mechanismus Link Fragmentation & Interleaving

- Low-Latency Queuing (LLQ) – Jedná se v podstatě o WFQ s jednou další prioritní frontou. Během cyklu výběru z front pomocí WFQ se vždy mezi výběry z jednotlivých front obsluhovaných VFQ nahlédne do prioritní fronty a odebere z ní jeden paket.

Existují i složitější kombinované režimy obsluhy. Například metoda Class-Based Weighted Fair Queuing (CBWFQ) je vlastně kombinací Custom Queueing a Weighted Fair Queuing. Při této metodě jsou jednotlivé kategorie provozu rozdělovány do několika front, z nichž každé je přiřazena určitá část kapacity výstupní linky. Pakety explicitně nezařazené do žádné z front jsou ukládány do implicitní fronty, v rámci níž jsou vyhledávány datové toky a těm je ze zbývající kapacity výstupní linky přidělováno pásmo metodou WFQ.

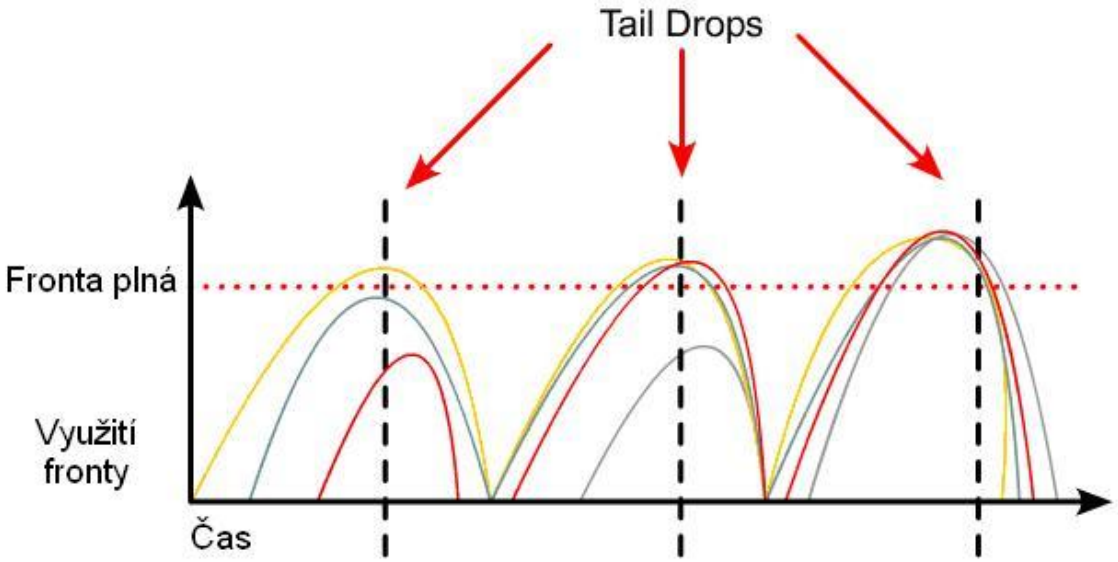
Ne všechny mechanismy podporující zajištění kvality služby musí být realizovány pouze režimem obsluhy front. Poměrně častým problémem je zajištění malého zpoždění pro interaktivní aplikace s krátkými pakety, pokud jsou tyto provozovány na pomalých linkách (typicky pod 800 kbps), na nichž je významné serializační zpoždění a po kterých rovněž prochází datový provoz s dlouhými pakety. Aby nemohlo dojít k situaci, že vysílání dlouhého datového paketu obsadí linku na tak dlouhou dobu, že nebudou moci být vysílány krátké pakety interaktivní aplikace (např. IP telefonie) s požadovaným maximálním rozestupem, mohou být dlouhé pakety před zahájením vysílání fragmentovány a jednotlivé fragmenty proloženy pakety aplikací citlivých na zpoždění. Uvedený mechanismus bývá označován jako Link Fragmentation & Interleaving (LFI).

1.4.1.4.3 Mechanismy předcházení zahlcení

V předchozí kapitole jsme popsali, jakým způsobem mohou síťové prvky řešit zahlcení (congestion management). Existují však také mechanismy, jak zahlcení předcházet. Předcházení zahlcení (congestion avoidance) je založeno na zpomalování zdroje dat na principu ovlivňování zpětné vazby od přijímače. Zpomalovat tak lze samozřejmě jen toky, které takovouto zpětnou vazbu implementují. Přímo na transportní vrstvě je zpětná vazba vestavěna do protokolu TCP, který je v současné době používá převládající množství síťových aplikací.

Problémem, který v sítích bez správné implementace mechanismů QoS často nastává, je globální synchronizace TCP spojení. Jestliže totiž dojde ve směrovači k úplnému zaplnění výstupní fronty některé linky, jsou všechny další pakety určené pro tuto linku zahazovány (mluvíme o tzv. tail drop, jelikož se zahazují pakety na konci fronty, resp. za ním). Pokud přes zmíněnou linku prochází větší množství TCP spojení, budou ze všech těchto spojení zahazovány pakety, takže nedorazí na přijímač, čímž ani vysílač neobdrží několik následných potvrzení. Na základě toho vysílač dojde k závěru, že síť je momentálně zahlcena, klesne s intenzitou vysílání na nulu a a aplikuje pomalý exponenciální náběh intenzity vysílání (proceduru Slow start). Nepříjemné však je, že proceduru Slow start zahájí téměř ve stejném okamžiku všechny toky, které byly přerušeny následkem likvidace jejich paketů směrovačem z důvodu nedostatečné kapacity fronty. Zatížení v síti tak začne oscilovat mezi nulovým (všechny toky současně zahájily Slow start) a maximem, při kterém opět dojde k přeplnění front, zahazování paketů všech toků a opětovnému pomalému náběhu současně u všech toků. Podstata problému tkví v tzv. globální synchronizaci TCP spojení, kdy nejsou pomalé náběhy jednotlivých TCP spojení náhodně rozložené v čase, čímž by sumární intenzita toku na lince

vedla k jisté průměrné hodnotě, ale vzájemně synchronizované a vytvářející pilovitý průběh zatížení sítě (obr. 1.23).



Obr. 1.23 - Globální synchronizace TCP spojení

Globální synchronizaci lze efektivně předejít aplikaci mechanizmu nazývaného Random Early Discard (RED), resp. Weighted Random Early Discard (WRED). Jeho princip spočívá v tom, že při zaplnění fronty nad určitou hranici se začnou se vzrůstající pravděpodobností (exponenciálně úměrnou zaplnění fronty) zahazovat pakety. Tím se jedno z TCP spojení, kterému zahozený paket patřil, zbrzdí dříve než ostatní, jejichž data dosud stále procházejí. Slow start všech spojení tak nenastává synchronně při úplném zaplnění fronty a zahazování všech dalších paketů přicházejících do fronty (tail drop), jelikož některé spojení zahájí vlivem zásahu algoritmu WRED proceduru Slow start dříve, čímž celkové zahlcení linky klesne.

V praxi se RED častěji aplikuje ve váhované verzi WRED, kdy se při překročení hranice pro náhodné zahazování paketů přednostně zahazují pakety s nižší prioritou (DSCP). Pravděpodobnost zahození v tomto případě není jen funkcí míry naplnění fronty, ale i priority paketu.

Při aplikaci RED je třeba pamatovat na to, že tato metoda omezuje pouze toky nad TCP protokolem. Pro omezení toků nad protokolem UDP by bylo třeba ovlivňovat zpětnou vazbu na aplikační úrovni, pokud takovouto vazbu konkrétní aplikace vůbec implementuje.

1.4.1.5 Nasazování mechanismů řízení kvality služby

Nasazování mechanismů podpory QoS do počítačové sítě zpravidla probíhá v těchto krocích:

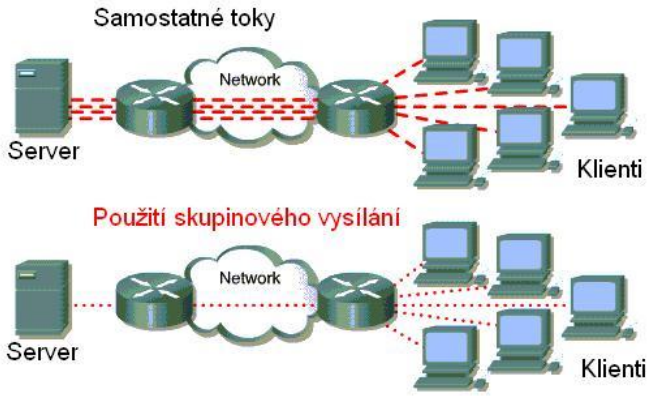
1. Zjištění provozovaných aplikací a jejich požadavků, monitorování stávajícího provozu
2. Vytvoření strategie podpory QoS a aplikace příslušných mechanismů QoS
3. Ověření chování mechanismů QoS. Zde se může ukázat, že výsledné chování sítě je z pohledu uživatelů horší, než před implementací. V této situaci je samozřejmě nutné strategii implementace QoS přehodnotit.

Úspěšnou implementací podpory QoS samozřejmě proces nekončí. Jelikož na živé síti postupně vyvstává potřeba provozu dalších a dalších aplikací, požadavky na QoS z pohledu uživatelů se budou neustále vyvíjet. Proto je třeba požadavky periodicky analyzovat a výše uvedené kroky vždy znovu opakovat.

1.4.2 Skupinové vysílání v sítích s protokolem IP

Pro implementaci multimediálních aplikací, při jejichž provozu je obvykle třeba distribuovat identický intenzivní tok dat k menší či velké skupině příjemců, je mimo podpory QoS užitečné v síti podporovat skupinové vysílání. Skupinové vysílání (angl. multicasting) znamená možnost vysílání pro určitou skupinu příjemců. Na rozdíl od všesměrového vysílání (broadcastingu), kdy je možno poslat paket všem příjemcům na lokální síti, však skupinové vysílání není omezeno hranicí lokální sítě. Příjemci skupinového vysílání mohou být rozprostřeni po rozlehlé síti nebo i po Internetu (resp. těch částech Internetu, které skupinové vysílání podporují).

Skupinové vysílání je výhodné použít všude tam, kde je třeba jednu informaci doručit skupině stanic, které mají o tuto danou informaci v jistém okamžiku zájem, aniž bychom museli jednotlivé proudy dat nesoucí požadovanou informaci od zdroje posílat nezávisle ke všem zájemcům. Zdroj generuje pouze jeden tok dat, které prvky síťové infrastruktury zasílají pouze do směrů, ve kterých leží zájemci o daný datový tok. Pro jednotlivé příjemce se pakety duplikují až v místech, kde se datový tok rozděluje do více větví sítě obsahujících příjemce multicastové skupiny (obr. 1.24). Spotřebované přenosové pásmo na linkách, za nimiž leží více příjemců, je tak výrazně menší, než kdyby jimi procházely toky generované zdrojem zvlášť pro jednotlivé příjemce.



Obr. 1.24 Srovnání zátěže při samostatných tocích a skupinovým vysílání

Způsob distribuce multicast paketů se poněkud liší v současných typických LAN a WAN sítích. V LAN je situace jednodušší že topologie neobsahují smyčky, takže není třeba řešit mechanismy vytváření logického stromu, po kterém se od zdroje budou multicast pakety šířit jen do větví sítě, kde je zájem o příjem příslušné skupiny.

1.4.2.1 Použití skupinového vysílání

Praktických aplikací, ve kterých je tentýž tok dat užitečné distribuovat skupině zájemců, je celá řada. Patří mezi ně zejména

- distribuce vysílání videa či hlasu (televize, rozhlas)
- videokonference mezi skupinou účastníků
- vyhledávání služeb určitého typu v rozlehlé síti
- distribuované simulace (včetně nejrůznějších typů internetových her)

1.4.2.2 Skupinové adresy na 2. a 3. vrstvě modelu OSI RM

Filosofii skupin při skupinovém vysílání lze pojmout různým způsobem. Skupinové vysílání v současných technologiích LAN i v protokolu IP využívá koncepce otevřených skupin. Otevřenost skupin znamená jednak to, že do skupiny se může přihlásit kterákoli stanice a jednak že do skupiny smí vysílat i stanice, která členem skupiny sama není. Stanice může být současně členem i více skupin. Přihlašování do skupiny je zcela v režii potenciálního přijímače každé skupiny, zdroj dat vysílající do skupiny nemůže žádným způsobem zjistit identitu ani počet příjemců, kteří v daném okamžiku data skupiny přijímají.

1.4.2.2.1 Skupinové MAC adresy

Na druhé vrstvě OSI RM používá většina technologií LAN k adresování stanic MAC adresy. Jejich rozsahy (první 3 bajty) jsou přidělovány jednotlivým výrobcům, kteří pak druhé tři bajty přidělují jednoznačně jednotlivým kusům jimi vyráběných síťových rozhraní. Mimo jednoznačně přidělených adres však MAC adresy mohou být i skupinové. IEEE definuje skupinové adresy jako ty, které mají první bit prvního bajtu MAC adresy vysílaného na médium nastaven na hodnotu 1 (zde si je dobré uvědomit, že u technologie Ethernet je to nejméně významný bit-LSB).
V souvislosti s protokolem IP se však jako skupinové MAC adresy využívají pouze adresy začínající na trojici bajtů 01-00-5E, jak bude vysvětleno v další kapitole.

Je důležité si uvědomit, že na 2. i 3. vrstvě mohou multicastové adresy být vždy jen adresami cíle - zdrojová adresa je vždy jednoznačná a identifikuje odesílatele dat. Této skutečnosti se využívá i při mnohých metodách distribuce dat skupinového vysílání k příjemcům ve WAN.

1.4.2.2.2 Skupinové adresy v IPv4 a IPv6, výběrové adresy

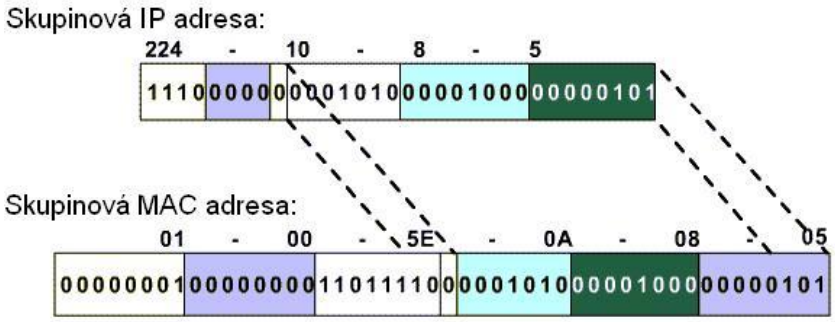
V protokolu IPv4 je pro skupinové adresy vyhrazen rozsah adres třídy D, tedy adresy 224.0.0.0 až 239.255.255.255. Rozsah adres 224.0.0.0 - 224.0.0.255 je rezervován pro služební (zejména směrovací) protokoly vyměňující si informace pouze v mezi sousedy na lokálním segmentu (TTL v hlavičce paketu je pro jistotu vždy nastaveno na 1). Pro aplikace univerzálního použití využívající multicastingu (např. NTP) přiděluje organizace IANA jednoznačné adresy z rozsahu 224.0.1.0 - 238.255.255.255. Adresy začínající trojicí bajtů 233.x.y mohou využívat bez dalších přidělovacích procedur správcové autonomního systému s číslem kódovaným do dvojice bajtů s hodnotami x a y (tzv. GLOP adresy). Existují také „privátní“ skupinové IP adresy s platností omezenou pouze na určitou oblast sítě 239.0.0.0 - 239.255.255.255, kterou nesmí opustit (tzv. Limited Scope Addresses, RFC 2365).

U protokolu IPv6 je skupinové vysílání využíváno ještě šířeji. Mimo otevřených skupin ve stylu IPv4 jsou zde definovány také tzv. výběrové adresy, umožňující vysílat nejen všem, ale jen „nejbližšímu“ členovi určité skupiny. Detaily lze najít v kapitole 1.2.2.1.

1.4.2.2.3 Mapování skupinových IP adres na MAC adresy

V koncových lokálních sítích musíme řešit mechanismus mapování skupinových IP adres na skupinové MAC adresy, podobně jako to pro individuální adresy řeší takovéto mapování protokol ARP. Mapování potřebujeme proto, abychom IP paket se skupinovou adresou mohli vložit do rámce a správně určit jeho cílovou MAC adresu.

Způsob mapování skupinových IPv4 adres na MAC adresy, který se v současné době ujal v praktických aplikacích se může zdát poněkud neefektivní, je to však dáno historickými důvody, za nichž v akademickém prostředí s omezenými prostředky na nákup rozsahů MAC adres mechanismus vznikal. Pro skupinové MAC adresy odpovídající skupinovým IPv4 adresám je vyhrazena polovina rozsahu daného prefixem 0x01-00-5E (všimněte si, že bit identifikující "multicast" v prvním bitu vysílaném na médium je nastaven). Z důvodu použití jen poloviny rozsahu adres začínajících na 0x01-00-5E je v nejvyšším bitu 4. bajtu MAC adresy vždy 0, takže pro mapování zbývá 23 bitů MAC adresy. Mapovaná IPv4 adresa má délku 32 bitů, ale protože všechny skupinové IP adresy patří do třídy D, začínají vždy kombinací bitů 1110, které není třeba mapovat. K mapování tedy zbývá 32-4=28bitů, které však musí být namapovány pouze na 23 bitů MAC adresy, jež jsou pro namapování k dispozici. Mapování se děje jednoduše tak, že se posledních 23 bitů skupinové IPv4 adresy přímo zkopíruje do posledních 23 bitů MAC adresy (obr. 1.25). Všimněme si, že 5 nejvyšších bitů skupinové IPv4 adresy se nemapuje, takže se vždy 32 multicast skupin protokolu IP mapuje na tutéž skupinovou MAC adresu. Například adresa 224.10.8.5 se mapuje na 01.00.5e.0a.08.05, stejně tak ale i další adresy vyjádřené binárním schematem 1110xxxx.x0001010.00001000.00000101, v němž hodnota bitu označeného symbolem x je libovolná.



Obr. 1.25 – Mapování skupinových adres IPv4 na MAC adresy

Nejednoznačnost mapování, při kterém síťové rozhraní spolu s požadovanou skupinou přijímá rámce i 31 dalších skupin, řeší až ovladače IP protokolu, které nežádoucí pakety odfiltrují. Proto je vhodné adresy používaných skupin volit tak, aby k překrývání prakticky nedocházeno.

U skupinových adres IPv6 je mapování poněkud jednodušší. Je pevně stanoven 2-bajtový prefix multicastových MAC adres a nejnižší 4 bajty IPv6 adresy se přímo kopírují do nejnižších 4 bajtů MAC adresy, jak je detailněji vysvětleno v kap. 1.2.2.1.

1.4.2.3 Skupinové vysílání v LAN

Skupinové vysílání v lokální síti je poměrně jednoduché, jelikož topologie lokální sítě neobsahuje smyčky. Výjimkou je kruhová topologie, zde však rámec vždy po oběhu kruhem odstraňuje stanice, která jej vyslala. Pokud jsou stanice připojeny na sdílené médium, rámec vyslaný pro skupinu stanic slyší všechny stanice. Je tedy na jednotlivých stanicích, zda rámec přijmou či nikoli. Pokud aplikace (resp. operační systém) na některé stanici rozhodne, že chce přijímat rámce určité skupiny, pouze instruuje hardware síťové karty, je má mimo rámců s její vlastní cílovou MAC adresou přijímat i rámce s MAC adresou některé konkrétní multicastové skupiny.

Poněkud složitější může být situace v síti obsahující přepínače. Díky použití protokolu Spanning Tree ani zde nikdy nevzniká smyčka, takže lze multicastové rámce stejně jako broadcast rámce bez nebezpečí kopírovat na všechny porty mimo příchozího, čímž se dostanou ke všem stanicím. Přesně tímto způsobem se k multicastům chovají levné přepínače. Dokonalejší přepínače však dokáží zohlednit fakt, že na rozdíl od broadcast rámců není třeba multicast pakety kopírovat všude, ale pouze do větví, ve kterých se vyskytují příjemci dané multicastové skupiny. To se mohou dozvědět buďto nahlížením do protokolu IGMP (viz dále), kterým si stanice vyžádávají od lokálního směrovače provoz multicastové skupiny (tzv. IGMP Snooping) nebo od speciálního (proprietárního) protokolu zavedeného mezi lokálním směrovačem a přepínačem, jímž směrovač přepínači sděluje, které MAC adresy u něj projevily zájem o příjem multicastové skupiny s adresou mapovanou na určitou skupinovou MAC adresu.

1.4.2.3.1 Protokol IGMP

Pokud se stanice na LAN chtějí přihlásit k příjmu provozu určité multicastové skupiny, použijí k tomu protokol IGMP (Internet Group Membership Protocol, RFC1112). Jeho prostřednictvím signalizují lokálnímu směrovači, že má požádat směrovače v páteři o rozšíření distribučního stromu, aby pokrýval segment s přijímačem. Podobně mohou stanice signalizovat protokolem IGMP i situaci, že chtějí příjmu provozu určité skupiny zanechat. Směrovač lokálního segmentu v tomto případě ještě dotazem ověří, zda na segmentu skutečně nezbývá žádná další stanice se zájmem o provoz skupiny a v negativním případě zasílání multicastů na lokální segment ukončí a do WAN zašle žádost o prořezání distribučního stromu. Jelikož přijímač multicastového provozu může být nekorektně vypnut, aniž by informoval pomocí IGMP lokální směrovač, ověřuje směrovač periodickými dotazy na segmentu, zda stanice mají o příjem své multicastové skupiny stále ještě zájem.

1.4.2.4 Skupinové vysílání ve WAN

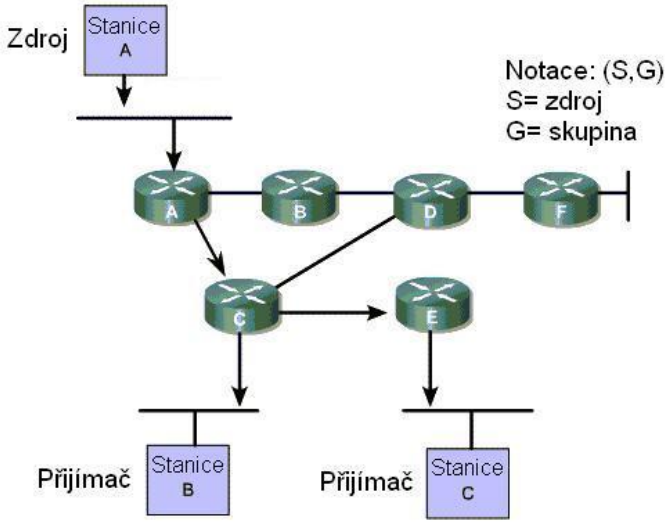
Distribuce multicast paketu ve WAN obsahující smyčky vyžaduje konstrukci tzv. distribučního stromu, jehož větve budou pokrývat všechny části WAN obsahující zájemce o příjem dané multicastové skupiny. Kořenem distribučního stromu je buďto směrovač sítě, kde je umístěn zdroj multicastů nebo dohodnutý směrovač na který zdroje multicastů tunelem posílají své pakety, pokud nechceme pro skupinu vytvářet tolik distribučních stromů, kolik je zdrojů skupinového vysílání.

Směrovač ve WAN síti musí na základě znalosti své pozice v distribučním stromu rozlišit, zda rozhraní, kterým multicast pakety příchází, vedou ke kořeni distribučního stromu a určit rozhraní, za kterými existují aktivní příjemci dané multicastové skupiny. Na základě informace o pozici směrovače v distribučním stromu pro danou skupinu si směrovače pro účely šíření multicastu budují zvláštní formu směrovací tabulky, obsahující informaci, které rozhraní vede ke zdroji multicastů a která do větví obsahující příjemce.

1.4.2.4.1 Distribuční stromy

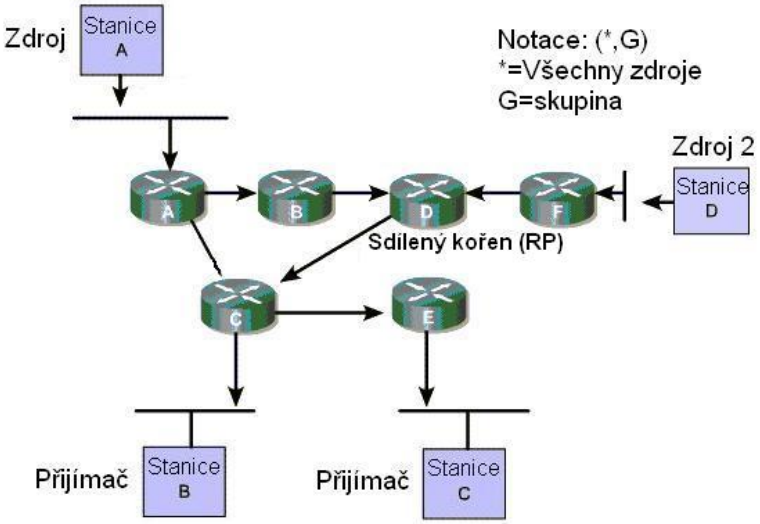
Distribuční strom tvoří acyklický podgraf grafu topologie sítě zahrnující všechny sítě, na nichž jsou přijímače dané multicastové skupiny. Jak se do skupiny přihlašují nové přijímače a odhlašují existující, musí se do stromu připojovat nové větve a odstraňovat větve neaktuální (mluvíme o „roubování“ a "prořezávání" stromu, angl. „grafting“, resp. "pruning").
Existuje několik přístupů, jak distribuční strom konstruovat. Podle toho, kam je umístěn kořen stromu rozlišujeme stromy na zdrojové (Source Tree, někdy též Shortest-Paths Tree) a sdílené (Shared Tree). Šíření provozu po stromě může být buďto jednosměrné (od kořene k listům) nebo obousměrné, při kterém se provoz šíří všemi směry od zdroje umístěného v některém (nekořenovém) uzlu stromu. Zdrojový strom jako strom nejkratších cest ze zdroje ke všem sítím s

přijímači skupiny představuje optimální, ale málo škálovatelné řešení. Provoz má sice nejmenší zpoždění, ale pro každý zdroj vysílání do každé multicastové skupiny musí existovat samostatný strom (ten se označuje < S,G >, kde S je zdrojová adresa, G je adresa multicastové skupiny). Příklad zdrojového stromu je vidět na obr. 1.26.



Obr. 1.26 – Zdrojový distribuční strom (vyznačen šipkami)

Sdílený strom je naproti tomu společný pro všechny zdroje ve skupině. Kořenem je vhodně zvolený směrovač (tzv. Rendezvous point, RP). Zdroje svá data zasílají na RP, ze kterého se pak šíří směrem dolů po distribučním stromě. Sdílené stromy se označují notací < *,G >, kde hvězdička informuje, že strom používají všechny zdroje přispívající do skupiny G. Pro mutlicastovou skupinu pak postačí jediný strom bez ohledu na počet přispívajících zdrojů. Nevýhodou je poněkud delší cesta paketů od zdroje k přijímačům přes RP a tím i větší zpoždění. Příklad zdrojového stromu ukazuje obr. 1.27.



Obr. 1.27– Sdílený distribuční strom (vyznačen šipkami)

Některé multicastové směrovací protokoly používají nejprve sdílený strom, který spotřebovává méně systémových zdrojů a až v případě příliš intenzivního provozu z některé zdrojové adresy (S) se zřídí pro danou multicastovou skupinu a tento zdroj < S,G > strom. Pokud existuje v multicastové směrovací tabulce záznam < S,G > i < *,G > a multicastový paket vyhoví oběma z nich, bude se směrování provádět podle specifičtějšího (< S,G >) záznamu po zdrojovém stromě.

1.4.2.4.2 Reverse Path Forwarding

Pro vytváření distribučního stromu byla postupně navržena celá řada mechanismů. Historicky prvním byl protokol DVMRP (Distance Vector Multicast Routing Protocol, RFC1075), který na obdobném principu jako RIP vyhledává nejkratší cesty ke všem zdrojům skupinového vysílání. Dále bylo navrženo rozšíření protokolu OSPF pro multicast (MOSPF) spočívající v tom, že jednotlivé směrovače formou LSA inzerují k nim připojené zájemce o jednotlivé multicastové skupiny a všechny směrovače na základě toho počítají strom nejkratších cest pokrývající přijímače jednotlivých skupin. V současné době se však téměř výhradně používá směrování na základě informace o zpětné cestě, tzv. Reverse Path Forwarding (RPF), které může být dále optimalizováno s použitím protokolu PIM (Protocol Independent Multicast).

Reverse Path Forwarding je metoda šíření multicast paketů od zdroje vysílání (resp. RP) po distribučním stromu. Je důležité si uvědomit, že při směrování multicastů se na rozdíl od běžného směrování směrovače orientují podle **zdrojové** adresy. Smyslem je paket šířit po distribučním stromu stále dál od zdroje vysílání (resp. RP), aniž by docházelo k jeho cyklování.

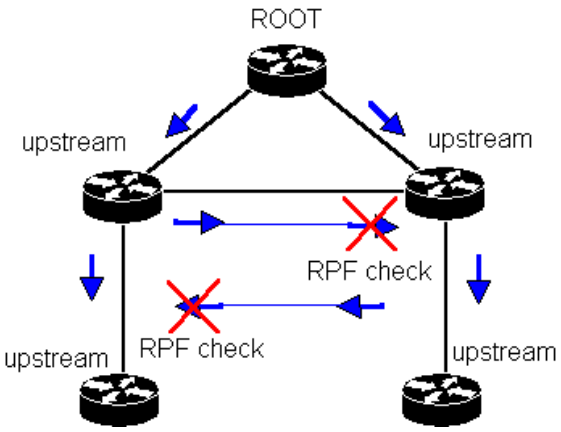
K určení rozhraní, kterými se má příchozí multicast paket dále rozesílat, slouží běžná (unicast) směrovací tabulka, vybudovaná jakýmkoli standardním směrovacím protokolem nebo i zadaná staticky. Proto hovoříme o směrování multicastů nezávislém na směrovacím protokolu (protocol independent multicast)

Z pohledu distribučního stromu pro danou multicastovou skupinu (a případně pro daný zdroj u < S,G > stromu) dělíme rozhraní každého směrovače na "downstream" a "upstream". Upstream rozhraní je právě jedno a vede směrem ke kořeni distribučního stromu. Je to rozhraní, kterým se vysílají běžné (unicast) pakety na adresu odpovídající kořeni distribučního stromu. Downstream rozhraní jsou rozhraní, která vedou do segmentů sítě, v nichž se vyskytují příjemci dané multicastové skupiny, ať už přímo stanice nebo další směrovače, které mají takové příjemce za svými downstream rozhraními.

Po rozdělení rozhraní směrovačů při směrování provozu určité skupiny na upstream a downstream můžeme vcelku jednoduše zajistit směrování provozu multicastové skupiny s vyloučením cyklení pomocí jednoduchého pravidla, nazývaného RPF Check, které můžeme formulovat takto:

Pokud multicast paket přišel z rozhraní, které se podle (unicast) směrovací tabulky používá pro směrování paketů ke zdrojové adrese multicastového paketu, resp. adrese RP (tedy upstream rozhraní), paket se rozešle na downstream rozhraní. V opačném případě se paket zahodí.

Pravidlo RPF Check zajistí rozšíření multicast paketu po celém distribučním stromě a vyloučí cyklování paketů ve smyčce. Pokud předpokládáme, že downstream rozhraní jsou všechna kromě jediného upstream rozhraní, zajistí se tím rozšíření paketu do celé sítě. Samotný RPF check však nedokáže zaručit, že multicast paket neprojde některou síti vícekrát, jak je ukázáno na obr. 1.28.



Obr. 1.28 – Duplikování části provozu při použití RPF check

Pro vyloučení tohoto případu je třeba mezi směrovači připojenými na tutéž síť zavést speciální protokol, kterým se mezi sebou dohodnou, který z nich bude na síť provoz multicastové skupiny šířit (bude to ten, který má k síti se zdrojem lepší metriku ve směrovací tabulce) . To je právě jedna z funkcí protokolu PIM. Dalšími funkcemi PIM jsou prořezávání větví, ve kterých nejsou žádní příjemci multicastové skupiny nebo naopak připojování větví, ve kterých se zájemci o skupinu objevili. Pomocí PIM se také směrovače na segmentu domlouvají, který z nich bude pomocí protokolu IGMP periodicky ověřovat, zda zájem příjemců o jistou multicastovou skupinu stále trvá (viz kap. 1.4.2.3.1). Řídící zprávy pro pořezání nebo připojení větve do distribučního stromu se vždy zasílají o úroveň distribučního stromu výše směrem ke kořeni (tj. zdroji multicast provozu nebo RP). Správný směr se určí pomocí standardní směrovací tabulky jako nejkratší cesta k síti se s adresou kořene distribučního stromu. Podle těchto zpráv se pro danou multicastovou skupinu na směrovačích rozšiřuje nebo naopak omezuje množina downstream rozhraní. Záznamy multicastové směrovací tabulky pak popisují aktuální stav distribučního stromu dané skupiny v okolí směrovače a pro zdrojové stromy budou mít tvar

< cílová multicast adresa, adresa zdroje multicastu,
upstream rozhraní,
množina downstream rozhraní >

Při použití zdrojového stromu musí směrovače na distribučním stromu udržovat tolik záznamů, kolik je vysílačů v multicastové skupině. Pro sdílené stromy se uchovává pro celou skupinu pouze jediný záznam, který bude mít tvar

< cílová multicast adresa, upstream rozhraní, množina downstream rozhraní >

Směrovače mimo distribuční strom dané skupiny nemusí o multicastové skupině uchovávat žádnou informaci.

Ke konstrukci distribučních stromů lze principiálně přistoupit dvěma způsoby. Jeden z nich bývá označován jako hustý režim (Dense mode), druhý jako řídký (Sparse Mode). V hustém režimu se předpokládá, že zájemci o multicastovou skupinu jsou skoro na všech segmentech sítě. Proto je provoz implicitně směrován do všech segmentů (všechna rozhraní směrovačů mimo upstream jsou na počátku považována za downstream). Pokud směrovač při příjmu multicastového paketu zjistí, že na žádném svém rozhraní nemá zájemce o multicastovou skupinu, pošle směrovači ve vyšší úrovni distribučního stromu žádost o "prořezání" (prune). Směrovač ve vyšší úrovni stromu na jejím základě přestane do větve multicastový provoz dané skupiny vysílat. Pokud se ve větvi časem objeví zájemce o multicastovou skupinu, může lokální směrovač segmentu větev do distribučního stromu opětovně přihlásit.

V řídkém režimu nejsou naproti tomu multicast pakety implicitně zasílány nikam. Většina větví by totiž z důvodu řídkého výskytu příjemců byla multicasty zaplavována zbytečně. Až když se na segmentu objeví zájemce o danou skupinu, pošle se směrem ke kořeni žádost o připojení nové větve distribučního stromu. Po této větvi pak k novému zájemci poteče provoz požadované multicastové skupiny.

Připojení větve do distribučního stromu v řídkém režimu nebo její prořezání v režimu hustém není trvalé. Každý směrovač distribučního stromu si totiž žádost o připojení, resp. prořezání pamatuje pouze po jistou omezenou dobu a po jejím vypršení přejde k původnímu chování. Proto se musí žádost o připojení a prořezání větve distribučního stromu periodicky opakovat.

Žádosti o připojení a prořezání větve stromu se mezi směrovači zpravidla zasílají prostřednictvím zpráv protokolu PIM. Z hlediska směrovače tak není rozdílu, jestli mulicastový provoz na rozhraní vyžaduje na segmentu připojená stanice nebo jiný směrovač, který multicastový provoz požaduje pro segmenty na jeho rozhraních.

V řídkém režimu se na řízení směrování účastní méně směšovačů - pouze ty, které vytvářejí obvykle ne příliš rozvětvený distribuční strom. Proto jsou protokoly pracující v řídkém režimu vhodnější pro rozlehlé WAN sítě.

Protokol PIM umí podporovat jak řídký, tak hustý režim – hovoříme pak o protokolu PIM v hustém (PIM-DM) nebo řídkém (PIM-SM) režimu. V současné době se v Internetu používá téměř výhradně protokolu PIM-SM.