The Role of Admission Control in Assuring Multiple Services Quality

Solange Rito Lima, Paulo Carvalho and Vasco Freitas
University of Minho, Department of Informatics, 4710-057 Braga, Portugal
{solange,pmc,vf}@di.uminho.pt

Abstract

Considering that network overprovisioning by itself is not always an attainable and everlasting solution, Admission Control (AC) mechanisms are recommended to keep network load controlled and assure the required service quality levels. This article debates the role of AC in multiservice IP networks, providing an overview and discussion of current and representative AC approaches, highlighting their main characteristics, pros and cons regarding the management of network services quality. In this debate, particular emphasis is given to an enhanced monitoring-based AC proposal for assuring multiple service levels in multiclass networks.

1. Introduction

To face today's Internet service heterogeneity and integration, the TCP/IP protocol suite has been enhanced with new service models, protocols and mechanisms. The Class of Service (CoS) paradigm, where flows with similar characteristics and service requirements are aggregated in the same class, has been pointed out as a suitable service model regarding scalable Quality of Service (QoS) support.

To control network resources efficiently and assure the required QoS levels, Admission Control (AC) has been recognized as a convenient traffic control mechanism [1, 2]. In fact, controlling the admission of flows entering the network and sharing a service class aims at avoiding overutilization of existing resources, satisfying the requirements of new incoming traffic flows without compromising the QoS of already active flows and, generically, preventing instability and congestion assuring QoS and SLSs fulfillment.

In general, the QoS guarantees and predictability required by a service class determines the control complexity inherent to an AC strategy. To obtain a good compromise between service guarantees, complexity and efficient resource utilization is a major challenge. Overprovisioning can be useful to improve this trade-off, however, a consistent QoS solution cannot just be based on overprovisioning and further control has to be in place to honor QoS requirements in the network. The challenge is increased when considering multiservice networks and end-to-end QoS delivery, as service classes have distinct characteristics requiring different QoS assurance levels, and multiple heterogeneous domains may be involved with negotiated SLSs' between them to be fulfilled.

The main objective of this document is to discuss existing AC proposals, covering their main characteristics, advantages and limitations in controlling multiple service levels. This analysis is relevant as an effective way to identify, understand and compare representative AC approaches, pointing out strategic directions for improving AC tasks. Facing this discussion, an enhanced AC proposal for managing QoS and SLSs in multiclass networks is presented.

The remaining of this document is organized as follows: relevant characteristics of existing AC approaches are identified and debated in Section 2; current and representative AC approaches are discussed in Section 3; the characteristics and key points of the AC model proposed for QoS and SLS control are highlighted in Section 4.

2. Relevant characteristics of AC approaches

Important high-level characteristics distinguishing AC approaches have been identified as follows:

(i) *the underlying network paradigm* - this aspect is related to the network model in which AC operates. AC approaches span from single service (best-effort) to multiservice architectures, following a flow or class-based paradigm. Their scope as regards targeting an intradomain, interdomain and/or end-to-end solution also varies;

(ii) *the type of service to control* - this aspect is closely related to the guarantee levels to be provided. Common and similar terminology includes guaranteed vs. predictive, guaranteed vs. controlled load or hard vs. soft real-time services. The type of service is tied up with the applications' characteristics, whether they are rigid or adaptive, have quantitative or qualitative QoS targets;

(iii) *the signaling support involved* - this topic can be viewed in two distinct ways. On the one hand, it is related to the type of applications and their ability to explicitly inform the network of their needs. This is commonly expressed in terms of a traffic profile and/or QoS requirements, using soft or hard state signaling for that purpose. On the other hand, signaling may also occur at high-level, for instance between specific nodes in distinct network domains or directly between end-systems. The nodes involved in the signaling process are closely related to the next topic;

(iv) *the location of the AC decision* - this aspect is related to the centralized or distributed nature of AC. This can be further detailed depending on which nodes (e.g., all nodes or specific nodes) are involved and how they participate in the AC process. For instance, a node can make an AC decision or only gather information for some other entity to use. The amount and type (per-flow or per-class) of state information kept in those nodes and the need for coordination among them are also important factors to consider;

(v) *the characteristics of the admission decision criteria* - these can be determined by (i) the nature of the algorithm, i.e., whether it is parameter-based, measurement-based or hybrid[1]; (ii) the information used for AC, which can be based on keeping track of resources' usage (usually bandwidth) or on congestion indicators (e.g., explicit congestion marks (ECN)); (iii) the concrete AC equations, which can be based on more or less intricate theoretical concepts involving distinct control parameters, whose tuning will, in turn, influence the conservativeness of AC.

Having discussed these points, the overall performance of an AC approach can be characterized through several related aspects, namely: (i) the ability to fulfill the QoS commitments; (ii) the efficiency of resources' utilization for the service levels provided; (iii) the overhead introduced in the network data and control planes influencing scalability; (iv) the latency regarding the time it takes to make an AC decision. The easy of migration and implementation in real environments is another key point as it brings a practical perspective and the real usefulness of the AC approach.

3. Detailing existing AC approaches

3.1. Intserv and RSVP aggregation

Although independent from the Intserv architecture [8], RSVP [9] is there pointed out as a convenient explicit setup mechanism to signal per-flow resource requirements in order to sustain node-by-node AC and resource reservation aiming at a guaranteed end-to-end QoS delivery [37][2].

The impairments of deploying Intserv/RSVP in large scale [5, 6] have motivated the aggregation of individual flow requests [3]. This aggregation process aims at reducing scalability problems, avoiding per-flow signaling and per-flow state information in the core, at cost of reducing the isolation among flows. In this process, interior nodes only maintain a reservation state for aggregates, and their state only changes when the corresponding aggregate reservation needs to be updated (increased or reduced) with a new bandwidth bulk.

The level of aggregation or bulk size influences the flows' admittance, the utilization and the demand for signaling in the core. While large bulks influence flow's acceptance and utilization negatively, small bulks influence these aspects positively at expense of more signaling. The need for signaling the aggregation region also depends on the traffic load variability and, ultimately, under high variability and low aggregation the process tends to per-flow reservation [16].

3.2. Intserv/Diffserv integrated solutions

According to the framework for Intserv/Diffserv operation [6], Intserv, RSVP and Diffserv are complementary technologies which can facilitate pursuing the objective of a scalable end-to-end quantitative QoS solution. While Intserv/RSVP allows per-flow request signaling quantifying the resources needed and obtaining a corresponding AC feedback, Diffserv enables scalability in large networks. In this framework, end-to-end RSVP signaling requires at least that RSVP messages are carried out across the Diffserv region, but depending on the specific realization of the framework, none, some (e.g. border router) or all routers in the Diffserv region, may process these messages. The coexistence between the two architectures assumes the control of the amount of traffic submitted to the Diffserv region, which must be able to support Intserv-like services through proper Per-hop Behaviors (PHBs), and Intserv/Diffserv parameters mapping (see Figure 1). The option for resource management in Diffserv region may include: (i) static provisioning; (ii) dynamic provisioning using RSVP; (iii) dynamic provisioning resorting to other means, such as Bandwidth Brokers (see Section 3.4).

In practice, despite the guarantees provided in the Intserv/RSVP region, with the inherent control overhead, end-to-end services guarantees depend on the resource management policies and supported services within Diffserv regions. A consistent mapping of Intserv/Diffserv services

[1] *Parameter-based* AC algorithms take into account the network resources already in use (reserved) by accepted flows and the resources the new flow will consume, according to its explicit traffic descriptor. *Measurement-based* AC algorithms take into account measures reflecting the impact of existing flows on the network load and/or QoS before deciding about a new admission.

[2] Recently, a comparison of existing QoS signaling protocols has been carried out in [23, 29] and the framework Next Steps in Signaling (NSIS) has been proposed [21]. This framework contemplates a wider variety of possible signaling scenarios, being more versatile and flexible than RSVP.

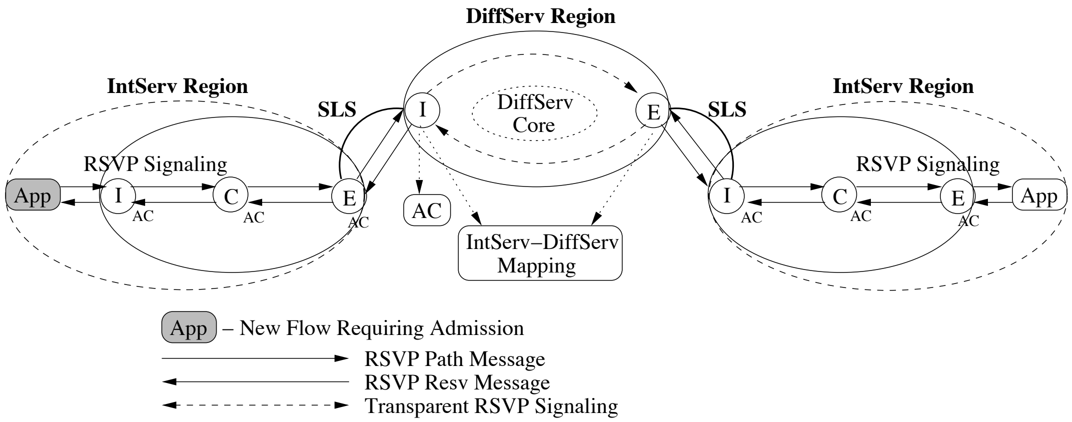

Figure 1. Intserv/Diffserv integrated solution

and parameters, an effective AC control to Diffserv regions and an effective control of resources inside this region to meet the services levels is essential to achieve end-to-end QoS guarantees.

3.3. DPS and SCORE architecture

The service architecture proposed in [34] aims at offering per-flow delay and bandwidth guarantees similar to the Intserv guaranteed service [8], but in a more scalable way. This architecture called Scalable Core (SCORE) is based on: (i) bringing per-flow management to edge nodes; (ii) a stateless core, i.e., a core where no per-flow information is maintained; (iii) a dynamic packet state (DPS) technique, which uses specific fields of the IP packet header to embed per-flow state. Core nodes process each packet based on its state, update it and, eventually, their own state before forwarding the packet.

DPS technique is the key concept of SCORE architecture allowing to coordinate routers' actions and implement distributed algorithms. The packet state inserted at ingress nodes and removed at egress nodes is used by each core node to perform scheduling (based on the concept of packet eligible time and deadline) and to support per-hop AC.

Despite avoiding per-flow state in the core and providing a guaranteed service, this architecture requires that all routers in the flow's path participate in the AC process, implement the same scheduling mechanism and update packet headers. The proposal for packet state insertion in packet headers may reveal itself incompatible with existing protocols and mechanisms such as Diffserv marking, headers compression and encryption. Although presented as an end-to-end solution, the operation crossing multiple domains is not covered in [34].

Within the Diffserv context, the framework Resource Management for Diffserv proposed in [36], although having a broader scope than SCORE, involves also two distinct signaling protocols: one acting on a per-hop basis called Per-Hop Reservation and other acting only at edge nodes called Per-Domain Reservation. At the edge nodes information is maintained per-flow. The information state at other nodes is PHB-based instead of flow-based, and AC can use either explicit reservations or measurements of traffic aggregates.

3.4. Centralized approaches based on BBs

One of the first approaches to perform resource management and AC in a Diffserv domain suggests the use of a central entity called Bandwidth Broker (BB) [32]. The principle behind BB architecture is to introduce in a Diffserv domain several service management tasks required to provide a consistent QoS, without complicating the control plane inside the network. This is achieved by centralizing information concerning network resources and their usage, domain topology, service policies, negotiated SLSs, which is required to perform control tasks such as AC, removing these tasks and the corresponding state information from the network core.

As far as AC is concerned, at an intradomain level, when a new flow requires admission, a signaling message is sent to the BB specifying the flow profile and QoS requirements. The BB, after authenticating and authorizing the request, makes a decision considering the domain service policies, the corresponding SLS usage and the available resources along the path. If the destination is outside the domain, the AC decision may involve interdomain signaling with downstream BBs, extending the AC process and resource reservation end-to-end. According to the resulting AC decision, each BB updates its state information databases and configures the involved edge nodes consistently. For scalability reasons, the AC requests to BBs, the reservations and the interdomain communication should consider flow aggregation. A more detailed description of the BB functionality and operation is available in [32, 35].

In [39], a BB's architecture based on a core stateless Virtual Time Reference System [38] and DPS technique is suggested to achieve a scalable solution to provide guaranteed services without requiring per-flow state in core routers. For Diffserv environments, a BB approach based on an Active Resource Management mechanism that reallocates dynamically bandwidth among clients has been proposed in [28].

The main advantage of centralized AC approaches is that centralizing state information and control tasks allows a global vision of the domain's QoS and operation, relieving the control plane inside the network. Centralization also facilitates creating and changing service policies and control mechanisms such as AC algorithms. The cost of centralized approaches is however high. BBs need to store and manage large amounts of information, which in large and highly dynamic networks with many signaling messages and information state updates needing to be processed in real-time is even hard or prohibitive. The congestion and functional dependence on a single entity is another well-known problem of centralization.

To improve reliability and scalability in large network domains, several approaches consider the use of a distributed or hierarchical architecture involving multiple BBs in the domain instead of a single centralized BB [13,32,39]. A single BB strategy is considered more suitable to small and less dynamic environments involving long lived flows. In the case of large and more dynamic domains, the use of multiple BBs improves reliability, BB congestion avoidance and scalability, at an eventual cost in coordination among BBs and in resources' fragmentation.

3.5. Measurement-based AC approaches

AC approaches based on network measurements performed node-by-node, edge-to-edge or end-to-end have erupted within the context of providing predictive service guarantees. They intend to solve or reduce the disadvantages of the described AC approaches, in particular, regarding the state information and control overhead, at an eventual cost of QoS degradation. Measuring network utilization and congestion can be expressed by the estimation and control of parameters such as bandwidth, delay, loss or ECN marks, during a given measurement period.

Passive measurement-based AC - The term *passive* stems from the fact that the measurement process resorts to real traffic within the network for parameters' estimation.

In the context of Intserv, MBAC has been proposed to assist the provision of a predictive service for tolerant applications able to accommodate occasional delay bound violations [22]. As the behavior of existing flows is determined by measurements rather than by their rate reservations (e.g., worst-case parameters), the service provided is less reliable due to traffic fluctuations. However, this allows to improve AC flexibility and to take advantage of statistical multiplexing, which may lead to significant utilization gains. In [22], AC is distributed node-by-node, using rate and/or delay-based equations. Other relevant MBAC algorithms are presented in [10, 18, 20].

Taking into account the burden of performing AC in all network nodes regarding the changes and overhead introduced in those nodes, a different type of passive MBAC considers measuring the edge-to-edge network status without requiring additional processing in the network core. AC is then left for network edges such as ingress nodes, egress nodes or both.

In the context of multiclass networks, [12,33] propose an AC solution based on the theory of traffic envelopes [33]. In this proposal, egress nodes assume a preponderant role as they perform both edge-to-edge aggregate traffic measurements and AC. The measurements assess passive and continuously the available service on a path between ingress-egress pairs, without involving per-flow state in any network node and ignoring core details.

Despite the scalability resulting from not involving the network core, the need for ingress-egress continuous measurements and updates in all real packets makes the solution more oriented to a single domain than to end-to-end. Moreover, the problematic of controlling SLSs is neither covered in this approach nor in the active AC proposals discussed below.

Active measurement-based AC - In opposition to passive measurement, the designation *active* measurement is adopted when specific traffic, called probing traffic, is injected into the network for measurement purposes. As regards AC, this technique intends to overcome the overhead associated with signaling and AC processing in network nodes, leaving uniquely to endpoints (end-systems or edge routers) the responsibility of inferring the network congestion status between them and of deciding on flow admission. This inference process resorts to per-flow probing traffic to obtain measures of delay, jitter, loss or ECN marks reflecting the congestion along the corresponding path, assessing simultaneously the path ability to support the new flow. AC approaches based on this technique are commonly called Endpoint Admission Control, Probe-based Admission Control or End-to-end Measurement-based Admission Control (EMBAC) [4, 7, 14, 19, 24]. Generically, in EMBAC, the admission of a new flow is preceded by a *probing phase* for congestion inference. The sender endpoint upon receiving AC feedback either enters in a *data phase*, where flow packets are sent, or aborts the sending process. In order to increase robustness, the sender implements a timeout mechanism associated with the start of the probing phase to deal with missing feedback. Figure 2 illustrates this behavior.

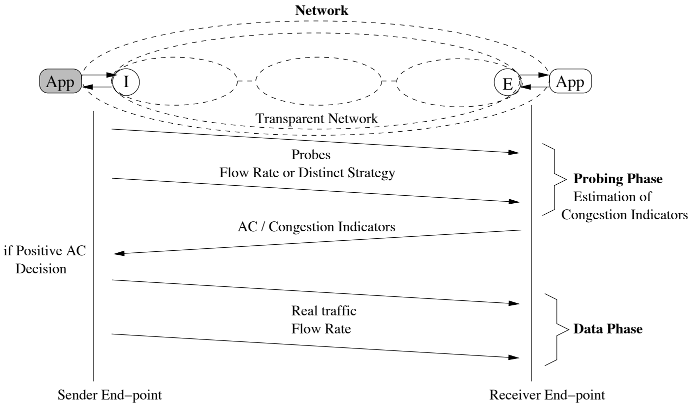

Figure 2. EMBAC approach

Existing EMBAC approaches differ in several aspects: (i) the measured parameter involved in the AC decision, e.g., packet loss ratio; (ii) the characteristics of the probing phase such as its duration and/or rate; (iii) the underlying network model. As regards the type of service to be applied to, EMBAC solutions are intended to have the same applicability of other measurement-based AC solutions, i.e., soft real-time services. A detailed discussion of EMBAC performance for a simplified network model with two priority service levels is available in [11]. Despite, the simplicity and scalability of EMBAC solutions, requiring none or reduced changes from networks, several disadvantages are commonly pointed out, namely: (i) the significant initial latency or setup delay which may limit its attractiveness for certain applications; (ii) the overhead of per-flow probing traffic which, depending on the weight and overlapping degree of the current probing phases, may lead to bandwidth stealing and thrashing regimes [11]; and (iii) the measurements' dependency on instantaneous network congestion.

3.6. AC proposals to control elastic traffic

Although performing AC for real-time traffic is generically consensual, the need to control the admission of elastic TCP traffic is more arguable, dividing opinions. While some argue that once TCP is adaptive controlling the number of flows sharing the available bandwidth is unnecessary, others are in favor saying that controlling the overload is required in order to preserve an acceptable throughput per active flow, and thus, the QoS offered to users [4,15,30,31]. In fact, a minimum TCP bandwidth is required to achieve a minimal session level user utility [31] and the use of AC will assure that, avoiding wasting network resources on retransmissions and incomplete transfers [30].

Due to the large number of TCP flow arrivals and their eventual small duration, controlling individual flows using explicit signaling and reservations is impracticable, therefore, in general, a measurement-based AC approach for elastic traffic is proposed to assure that the solution is able to react and scale properly. Without per-flow signaling, the detection and acceptance/rejection of a new flow is made implicitly. Common implicit AC criteria [4, 15, 31] use the estimation of current load, available bandwidth or packet loss probability, comparing the obtained estimation with a pre-defined threshold, which may depend on an upper limit of admitted flows. These estimates can relate to a link or path, however, path estimations are preferable when considering AC only performed at ingress nodes. In [15], several proposals for path estimations are summarized.

Within implicit AC the simple discard of initial flow packets is usually enough to inform the source of a rejection decision, otherwise those packets will proceed. In more detail, possible solutions to support detection and corresponding AC decision are: (i) to intercept packets initializing the TCP connection (TCP SYN and/or SYN ACK) [31]; (ii) to maintain a list of accepted and active flows based on the corresponding flow identifiers [15]. While the former solution is easy to implement, the latter is more flexible but critical for high-speed interfaces due to its potential overhead.

Note that, implicit AC can be applied to other traffic than TCP, for instance, to UDP traffic from real-time applications that do not send explicit signaling to the network.

4. A Monitoring-based AC proposal for Scalable QoS and SLS Control

Despite the wide range of AC approaches proposed in the literature, from which the most representative have been discussed above, few studies deal with the management of multiple intradomain QoS levels and interdomain SLSs simultaneously, lacking in formalizing a generic model with concrete and flexible AC equations to be deployed in CoS networks.

In this context, the AC model proposed in [25] and highlighted in this section brings new insights to perform encompassing and lightweight AC in multiservice class-based environments. The proposed AC model aims to: (i) support multiple services with distinct assurance levels; (ii) control the QoS levels inside each domain and the existing SLSs between domains; (iii) operate intra and interdomain providing an unified end-to-end solution; (iv) be simple, flexible, efficient, scalable and easy to deploy in real environments.

Facing the debate on related work, several aspects were identified as relevant for pursuing these objectives namely, distributing control between edge nodes, relieving network core from control tasks, reducing state information and control overhead, sensing and adapting to network dynamics through measurements, supporting AC irrespectively of applications' ability to explicit requirements and signaling the network. In addition, although not covered in the studied AC approaches, a certain degree of overprovisioning is considered to achieve a simple and manageable multiservice AC solution. This degree, which is service-dependent, aims at simplifying the AC process while providing the required service level guarantees.

4.1. AC Model architecture

In [25, 26], admission decisions consider both the levels of QoS being offered for each service type and the corresponding SLSs utilization. Therefore, the model architecture lays on service definition, QoS/SLS monitoring and CoS traffic characterization to sustain the definition and operation of the AC decision criteria, interrelated as shown in Figure 3.

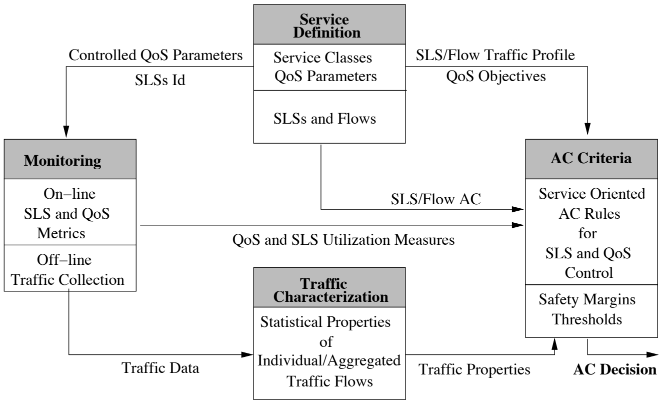

Figure 3. AC model architecture

Service definition defines services adjusted to different application requirements and the relevant QoS parameters to control within each service type. It also defines SLSs' syntax and semantics. Through systematic edge-to-edge measures of QoS parameters and SLSs utilization, *on-line monitoring* keeps track of QoS and SLS status in the domain through well-defined metrics, providing feedback to drive AC decisions. As an *off-line monitoring* process, CoS traffic aggregates may also be collected for subsequent off-line analysis and characterization. This analysis allows to determine the statistical properties of each class as a result of traffic aggregation so that more realistic service-oriented AC rules, thresholds and safety margins can be established. The knowledge resulting from interrelating these topics and from comparing existing measurement-based or hybrid AC algorithms provided the basics for defining a multiservice *AC decision criteria.*

4.2. Generic model operation

As illustrated in Figure 4, in the model operation only edge nodes are involved, leaving the network core unchanged. While ingress nodes perform explicit or implicit AC depending on the application type and corresponding traffic class, egress nodes perform on-line QoS and SLS monitoring. The *Ingress-Egress QoS Monitoring* task measures relevant parameters for each service (service metrics) using appropriate time scales and methodologies. The resulting measures are expected to reflect the service available from each ingress node, and are used by a QoS Control rule to drive AC decisions. This rule checks the controlled parameters of each service class against pre-defined thresholds to determine an AC status for the measurement time interval ($AC\_Status_{\Delta t_i}$). The *SLS Control* task monitors the usage of downstream SLSs at each egress, to ensure that traffic to other domains does not exceed the negotiated profiles and packet drop will not occur due to a simple and indiscriminate TC process. An SLS Rate Control rule, based on the Measure-Sum Algorithm, checks if the SLS can accommodate the traffic profile of the new flow, complementing the AC decision process. For implicit AC, as flows are unable to describe a rate profile, AC is restricted to the QoS control equation. Thus, flows are accepted or rejected implicitly according to the current $AC\_Status_{\Delta t_i}$, computed once for each measurement interval.

QoS monitoring statistics and SLS utilization are sent to the corresponding ingress routers to update an ingress-egress service matrix used for distributed AC and active service management. This notification may be carried out periodically, when a metric value or its variation exceeds a limit, or the SLS utilization exceeds a safety threshold.

The *end-to-end* operation is viewed as a repetitive and cumulative process of AC and available service computation[3] performed at ingress nodes. At each domain, the ingress node decides if a flow can be accepted and, if so, the service metric values in the domain are added to the flow request to inform the downstream domain of the service available so far. Using the incoming and its own measures each domain performs AC. More precisely, verifying if each flow's QoS parameter target value can be satisfied involves considering the corresponding QoS parameter bound in the domain and the cumulative value computed so far.

The performance evaluation of this AC model, reported in [26, 27], evinces the relevance and applicability of the defined AC rules, showing that the proposed solution is able to control multiple service levels efficiently.

[3] A cumulative process for end-to-end QoS computation is consistent with the cascade approach for the support of interoperator IP-based services, which is in conformance with the Internet structure and operation, and more scalable than the source-based approach [17].

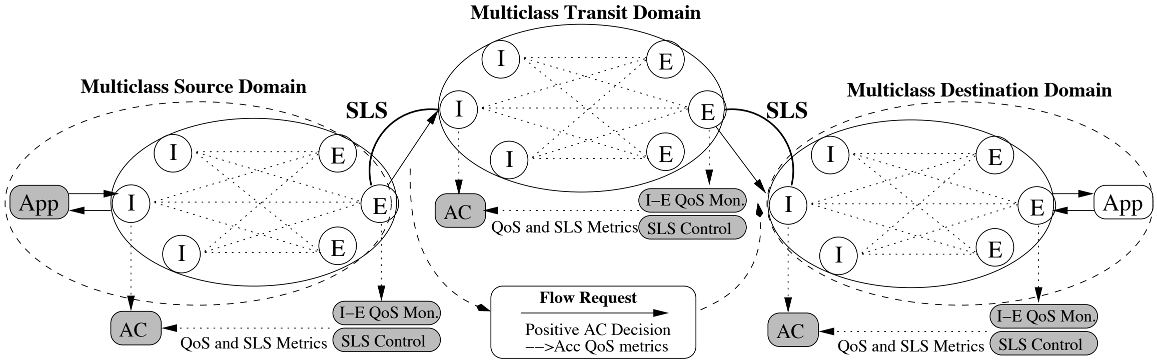

Figure 4. Distributed monitoring-based AC approach

4.3. AC model key points

The way service-dependent AC is proposed and articulated with on-line performance monitoring leads to important design aspects of the model, namely:

(i) different service types, QoS parameters and SLSs can be controlled simultaneously in a distributed and simple fashion, involving only edge nodes. This provides a convenient level of abstraction and independence from network core complexity and heterogeneity;

(ii) the state information is service and ingress-to-egress based which, apart from leading to reduced state information, is particularly suitable for SLS auditing;

(iii) the signaling process for intra and interdomain operation is simple, horizontal and fluid. The flow AC request is used both for per-domain AC and for end-to-end available service computation along the data path, and no soft/hard state behavior and symmetric routing paths are imposed;

(iv) the AC model provides enough flexibility to accommodate technological, service and application evolution. The service-dependent nature of AC rules and the conceptual modular independence between AC and monitoring tasks, increasing their ability to integrate new developments and improvements, contribute for the model's flexibility;

(v) the systematic use of on-line monitoring for traffic load and QoS metrics' estimation in a per-class basis, while allowing an adaptive service management, avoids per-application intrusive traffic to obtain measures and reduces AC latency as measures are available on-line. Furthermore, systematic measurements have an intrinsic auto-corrective nature, allowing to detect short or long-term traffic fluctuations depending on the measurement time interval, and implicitly take into account the effect of cross-traffic and other internally generated traffic (e.g., routing and management).

5. Conclusions

The need to perform AC in multiservice IP networks has been discussed and representative AC approaches surveyed facing the provision of multiple service levels. Conceptually, the present discussion clearly illustrates the compromise between the level of QoS guarantees and the complexity introduced in the network control plane. A broad view over the AC approaches evolution exhibits a tendency in adopting solutions based on measurements of network usage and performance rather than solutions bringing too much state information about reserved resources into the network. In this context, an encompassing and lightweight monitoring-based AC proposal for multiservice networks has been presented and its major key points regarding a scalable QoS and SLS control discussed.

References

[1] R. Atkinson, S. Floyd, and I. A. Board. IAB Concerns and Recommendations Regarding Internet Research and Evolution. RFC 3869 (Informational), Aug. 2004.

[2] J. Babiarz, K. Chan, and F. Baker. Configuration Guidelines for Diffserv Service Classes. IETF Draft : draft-ietf-tsvwg-diffserv-service-classes-02.txt (work in progress), Feb. 2006.

[3] F. Baker, C. Iturralde, F. L. Faucheur, and B. Davie. Aggregation of RSVP for IPv4 and IPv6 Reservations. RFC 3175 (Proposed Standard), Sept. 2001.

[4] N. Benameur, S. Fredj, F. Delcoigne, S. Oueslati-Boulahia, and J. Roberts. Integrated Admission Control for Streaming and Elastic Traffic. In M. Smirnov, J. Crowcroft, J. Roberts, and F. Boavida, editors, *QofIS'01*, volume 2156, pages 67–81, Sept. 2001.

[5] Y. Bernet. The complementary Roles of RSVP and Differentiated Services in the Full-Service QoS Network. *IEEE Communications Magazine*, pages 154–162, Feb. 2000.

[6] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine. A Framework for Integrated Services Operation over Diffserv Networks. RFC 2998 (Informational), Nov. 2000.

[7] G. Bianchi, A. Capone, and C. Petrioli. Throughput Analysis of End-to-End Measurement-based Admission Control in IP. In *IEEE INFOCOM'00*, 2000.

[8] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture: an Overview. RFC 1633 (Informational), June 1994.

[9] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. RFC 2205 (Proposed Standard), Sept. 1997. Updated by RFCs 2750, 3936.

[10] L. Breslau and S. Jamin. Comments on the Performance of Measurement-Based Admission Control Algorithms. In *IEEE INFOCOM'00*, Mar. 2000.

[11] L. Breslau, E. Knightly, S. Shenker, I. Stoica, and H. Zhang. Endpoint Admission Control: Architectural Issues and Performance. In *ACM SIGCOMM'00*, 2000.

[12] C. Cetinkaya, V. Kanodia, and E. Knightly. Scalable Services via Egress Admission Control. *IEEE Transactions on Multimedia*, 3(1):69–81, Mar. 2001.

[13] Z. Duan, Z. Zhang, Y. Hou, and L. Gao. A Core Stateless Bandwidth Broker Architecture for Scalable Support of Guaranteed Services. *IEEE Trans. Parallel Distrib. Syst.*, 15(2):167–182, 2004.

[14] V. Elek, G. Karlsson, and R. Rnngren. Admission Control Based on End-to-End Measurements. In *IEEE INFOCOM'00*, 2000.

[15] S. Fredj, S. Oueslati-Boulahia, and J. Roberts. Measurement-based Admission Control for Elastic Traffic. In *17th Int. Teletraffic Congress*, June 2001.

[16] H. Fu and E. Knightly. Aggregation and Scalable QoS: A Performance Study. In *IWQoS'01*, June 2001.

[17] P. Georgatsos, J. Spencer, D. Griffin, P. Damilatis, H. Asgari, J. Griem, G. Pavlou, and P. Morand. Provider-level Service Agreements for Inter-domain QoS delivery. *Fourth International Workshop on Advanced Internet Charging and QoS Technologies (ICQT04)*, Sept. 2004.

[18] R. Gibbens and F. Kelly. Measurement-based Connection Admission Control. In *15th International Teletraffic Congress*, June 1997.

[19] R. Gibbens and F. Kelly. Distributed Connection Acceptance Control for a Connectionless Network. In *16th International Teletraffic Congress*, June 1999.

[20] M. Grossglauser and D. Tse. A Time-Scale Decomposition Approach to Measurement-based Admission Control. *IEEE/ACM Trans. on Networking*, 11(4):550–563, Apr. 2003.

[21] R. Hancock, G. Karagiannis, J. Loughney, and S. V. den Bosch. Next Steps in Signaling (NSIS): Framework. RFC 4080 (Informational), June 2005.

[22] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A Measurement-Based Call Admission Control Algorithm for Integrated Services Packet Networks (Extended Version). *IEEE/ACM Trans. on Networking*, pages 56–70, Feb. 1997.

[23] P. Ji, Z. Ge, J. Kurose, and D. Towsley. A comparison of hard-state and soft-state signaling protocols. In *SIGCOMM '03*, pages 251–262, New York, NY, USA, 2003. ACM Press.

[24] F. Kelly, P. Key, and S. Zachary. Distributed Admission Control. *IEEE Journal on Selected Areas in Communications (JSAC)*, 18(12), Dec. 2000.

[25] S. R. Lima, P. Carvalho, and V. Freitas. Distributed Admission Control for QoS and SLS Management. *Journal of Network and Systems Management - Special Issue on Distributed Management*, 12(3):397–426, Sept. 2004.

[26] S. R. Lima, P. Carvalho, and V. Freitas. Self-adaptive Distributed Management of QoS and SLSs in Multiservice Networks. In *IEEE/IFIP Int. Conference on Integrated Management (IM 2005)*, Nice, France, May 2005. IEEE Press.

[27] S. R. Lima, P. Carvalho, and V. Freitas. Ensuring IP Services Consistency through Lightweight Monitoring-based Admission Control. In *11th International Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks*, Trento, Italy, June 2006. IEEE Press, ISBN:0-7803-9537.

[28] M. Mahajan, A. Ramanathan, and M. Parashar. Active Resource Management for the Differentiated Services Environment. *International Journal of Network Management*, 14(2):149–165, Mar. 2004.

[29] J. Manner and X. Fu. Analysis of Existing Quality-of-Service Signaling Protocols. RFC 4094 (Informational), May 2005.

[30] L. Massoulié and J. Roberts. Arguments in Favour of Admission Control for TCP Flows. In *16th International Teletraffic Congress*, pages 33–44, June 1999.

[31] R. Mortier, I. Pratt, C. Clark, and S. Crosby. Implicit Admission Control. *IEEE Journal on Selected Areas in Communication*, 18(12):2629–2639, Dec. 2000.

[32] K. Nichols, V. Jacobson, and L. Zhang. A Two-bit Differentiated Services Architecture for the Internet. RFC 2638 (Informational), July 1999.

[33] J. Qiu and E. Knightly. Measurement-Based Admission Control with Aggregate Traffic Envelopes. *IEEE/ACM Transactions on Networking*, 9(2):199–210, Apr. 2001.

[34] I. Stoica and H. Zhang. Providing Guaranteed Services Without Per Flow Management. In *ACM SIGCOMM'99*, Oct. 1999.

[35] B. Teitelbaum, S. Hares, L. Dunn, R. N. V. Narayan, and F. Reichmeyer. Internet2 QBone: building a testbed for differentiated services. *IEEE Network*, 13(5):8–16, "September/October" 1999.

[36] L. Westberg. Resource Management in Diffserv (RMD) Framework. IETF Draft: draft-westberg-rmd-framework-04.txt (working draft), Sept. 2003.

[37] J. Wroclawski. The Use of RSVP with IETF Integrated Services. RFC 2210 (Proposed Standard), Sept. 1997.

[38] Z. Zhang, Z. Duan, and Y. Hou. Virtual Time Reference System: A Unifying Scheduling Framework for Scalable Support of Guaranteed Services. *IEEE Journal on Selected Areas in Communication*, 18(12), Dec. 2000.

[39] Z. Zhang, Z. Duan, Y. Hou, and L. Gao. Decoupling QoS Control from Core Routers: A Novel Bandwidth Broker Architecture for Scalable Support of Guaranteed Services. In *ACM SIGCOMM'00*, 2000.