Decoupling QoS Control from Core Routers: A Novel Bandwidth Broker Architecture for Scalable Support of Guaranteed Services

Zhi-Li Zhang, Zhenhai Duan, Lixin Gao, and YiWei Thomas Hou

Virtual Time Reference System: A Unifying Scheduling Framework for Scalable Support of Guaranteed Services

Zhi-Li Zhang, Zhenhai Duan, and YiWei Thomas Hou

Speaker: Wei-Ming Yin

Instructor: Ying-Dar Lin

Nov. 24th, 2000

Agenda

- Motivation
- Virtual Time Reference System
  - Core stateless framework
  - End-to-end delay bound
- Bandwidth Broker Architecture
- Admission Control for Per-Flow Guaranteed Services
  - All rate-based vs. Mixed rate- and delay-based schedulers
- Admission Control for Class-Based Guaranteed Services
  - Dynamic flow aggregation under all rate-based schedulers
- Simulation Investigation
- Conclusion and Future works

Motivation

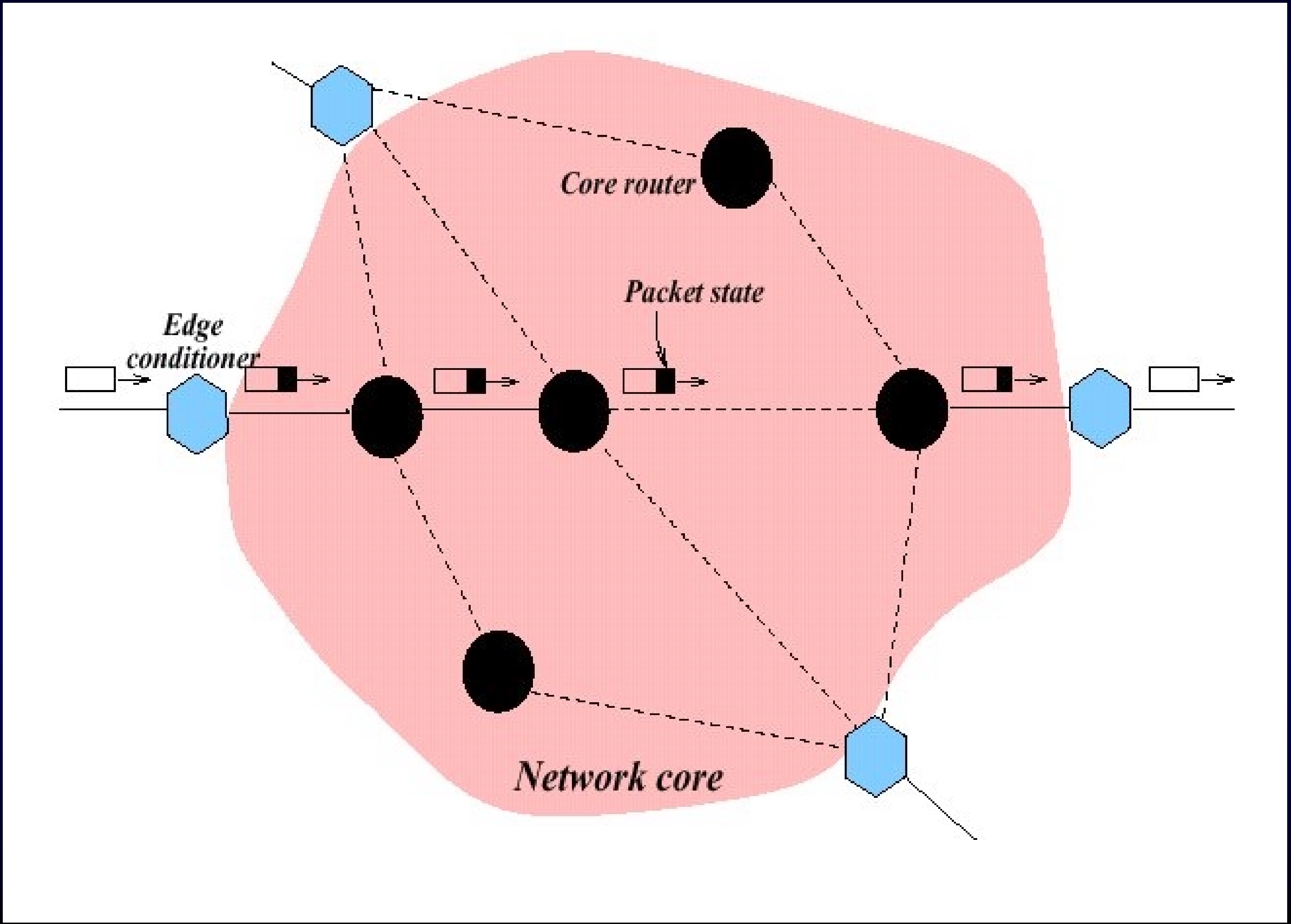
- Hop-by-hop admission control approach
  - Maintain per-flow or class-based QoS states at core routers
  - Perform local admission control and resource reservation
  - Maintain consistency of soft QoS states among all core routers
  - High communication overhead, less scalability, complicated design of core routers

- Path-oriented admission control approach
  - Relive core routers of QoS functions
  - Scale to both per-flow and class-based guaranteed services
  - Enable sophisticated QoS provisioning and admission control
  - No or minimal configuration of core routers

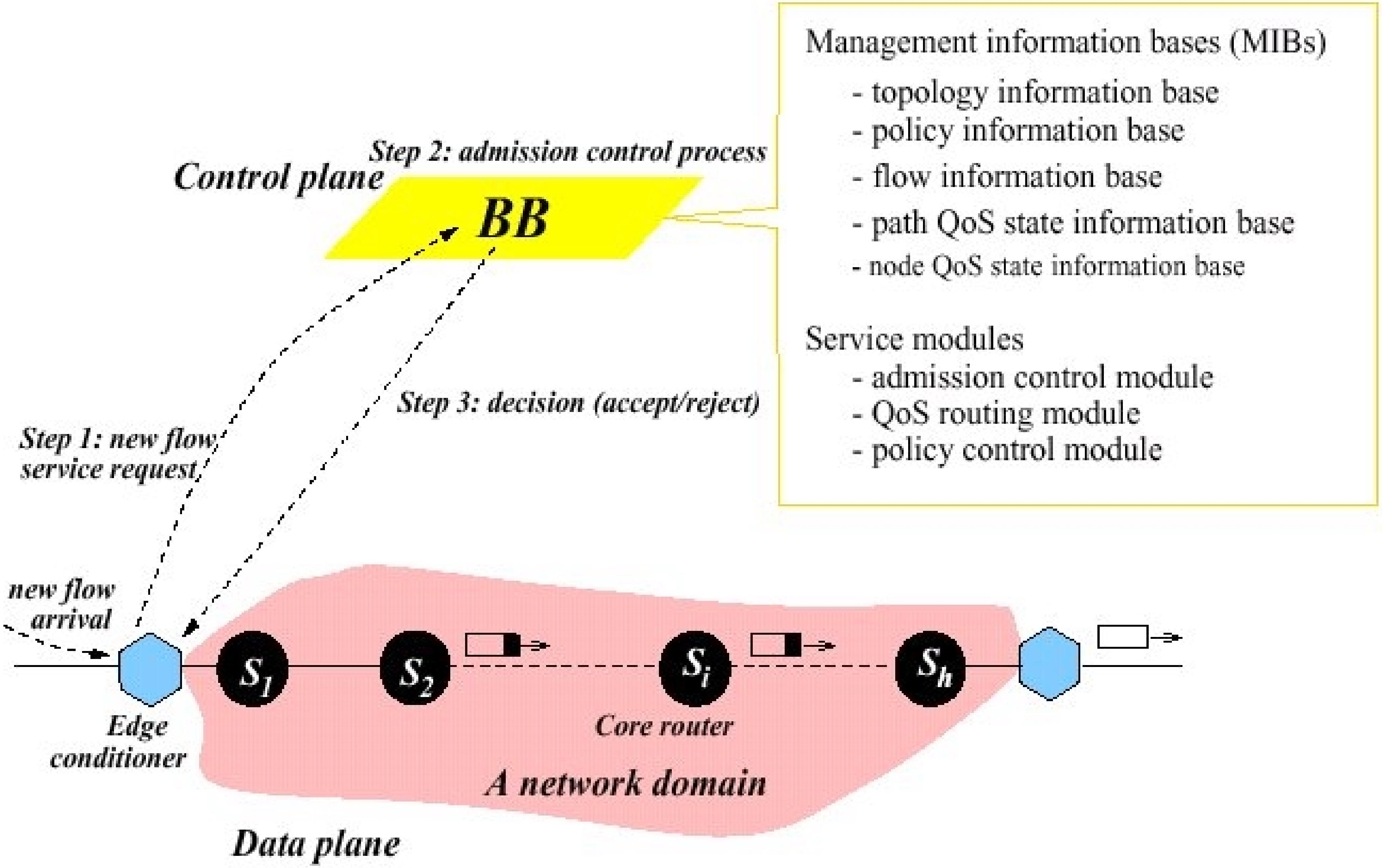
Virtual Time Reference System

- A core stateless framework
- A unifying scheduling framework
  - Core routers only perform forwarding and scheduling
- Three logic components
  - Packet state (on packet)
  - Edge traffic conditioning (edge)
  - Virtual time reference/update mechanism (core)
- Characterize per-hop behavior and end-to-end delay bound

Virtual Time Reference System



System Overview



Management information bases (MIBs)

- topology information base
- policy information base
- flow information base
- path QoS state information base
- node QoS state information base

Service modules

- admission control module
- QoS routing module
- policy control module

Dynamic packet state

- State types:

The kth packet of flow j at core router i.

The rate-delay parameter pair $\left(r^j, d^j\right)$. ← admission control

The virtual time stamp $w_i^{j,k}$. ← edge

The virtual time adjustment term $\delta^{j,k}$. ← edge

- Carried in packet header, initialized and inserted at edge, referenced (scheduling module) and updated (forwarding module) at core.

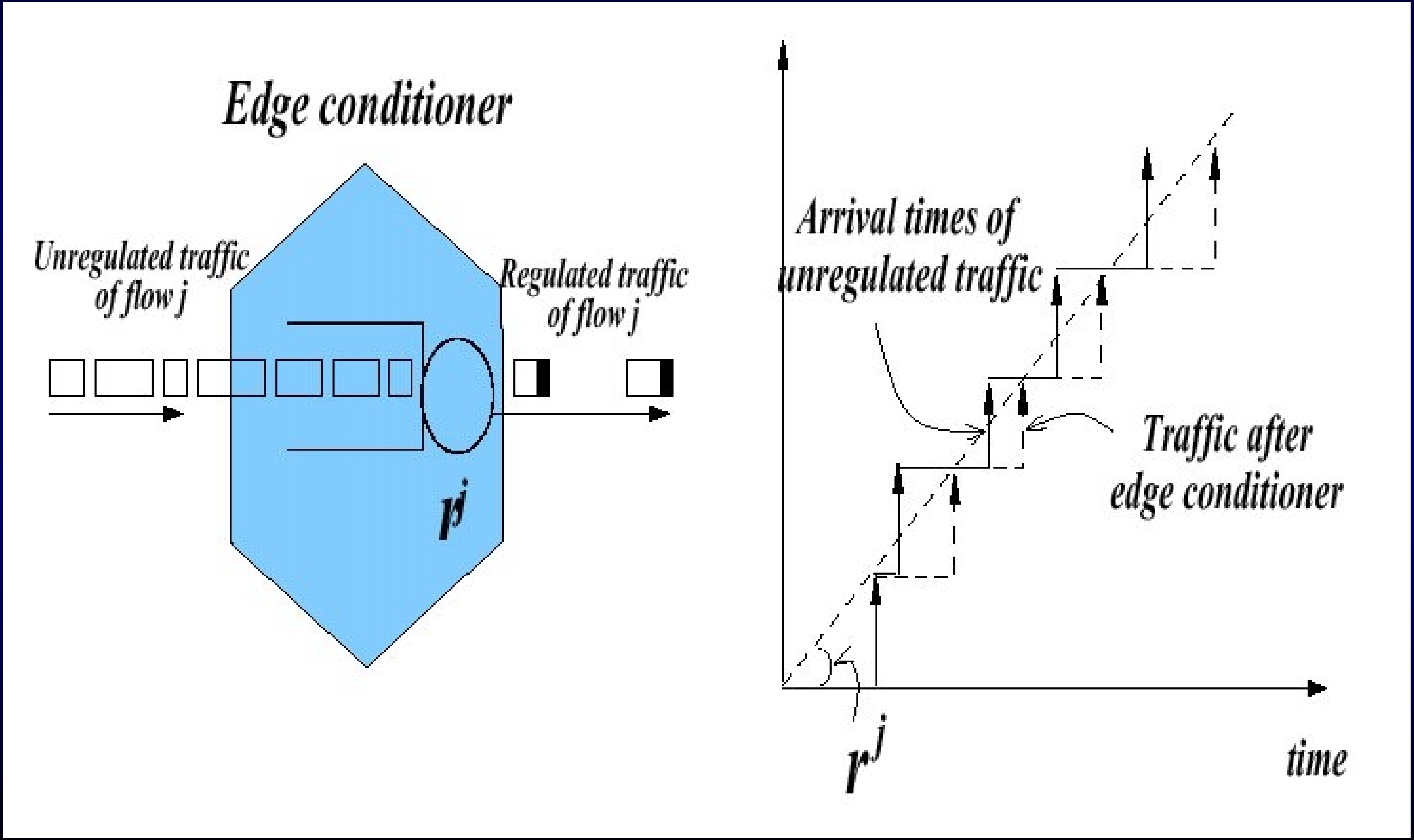
Edge Traffic Conditioning

- Regulate packets injection rate not exceeding reserved rate

$$a_1^{j,k+1} - a_1^{j,k} \geq \frac{L^{j,k+1}}{r^j},$$

where $a_1^{j,k}$ denotes the arrival time of kth packet of flow j, and $L^{j,k}$ denotes the size of that packet.

Edge Traffic Conditioning



Virtual Time Reference/Update

- Per-hop behavior

$$\text{virtual delay}: d_i^{j,k} = \begin{cases} \frac{L^{j,k}}{r^j} + \delta^{j,k}, & \text{rate - based scheduler} \\ d^j, & \text{delay - based scheduler} \end{cases}$$

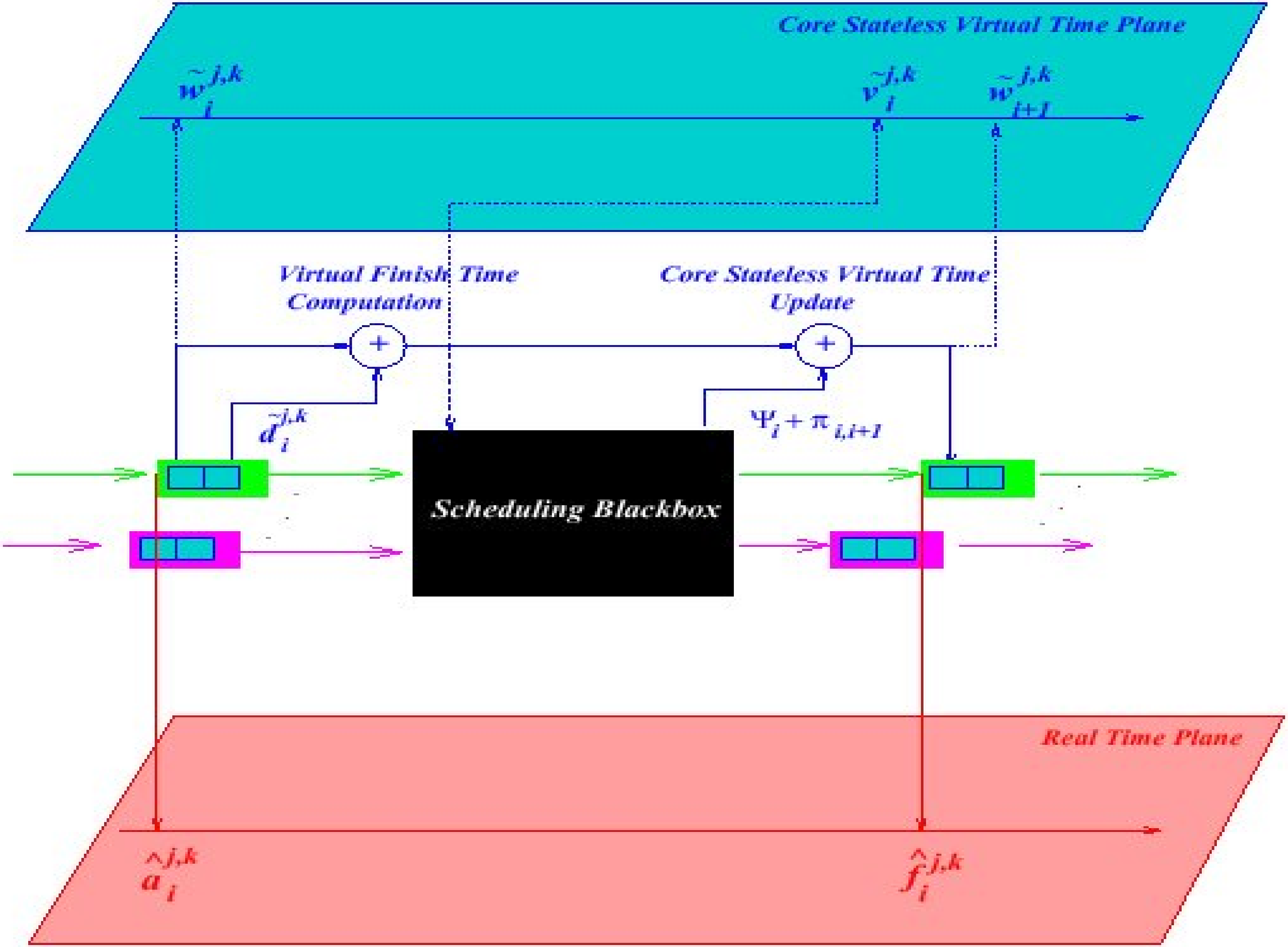
$$\text{virtual finish time}: v_i^{j,k} = w_i^{j,k} + d_i^{j,k} \quad \leftarrow \text{referenced}$$

$$\text{error term of core ronter } i : \Psi_{\mathrm{i}}$$

$$\text{actual finish time}: f_i^{j,k} \le v_i^{j,k} + \Psi_i \quad \leftarrow \text{per - hop behavior}$$

$$\text{propagation delay to next hop of core ruter } i : \pi_i$$

$$w_{i+1}^{j,k} = v_i^{j,k} + \Psi_i + \pi_i \quad \leftarrow \text{ updated}$$



Core Stateless Virtual Time Plane

Virtual Finish Time Computation

Core Stateless Virtual Time Update

Scheduling Blackbox

Real Time Plane

Virtual Time Reference System

- Suppose total *h* hops, of which *q* hop are rate-based scheduler, and *h-q* hops are delay-based schedulers. The traffic profile of flow *j* is (s^j, r^j, P^j, $L^{j,max}$).

$$f_h^{j,k} - a_1^{j,k} \le d_{core}^j = q\frac{L^{j,\max}}{r^j} + (h-q)d^j + \sum\nolimits_{i\in P}(\Psi_i + \pi_i)$$

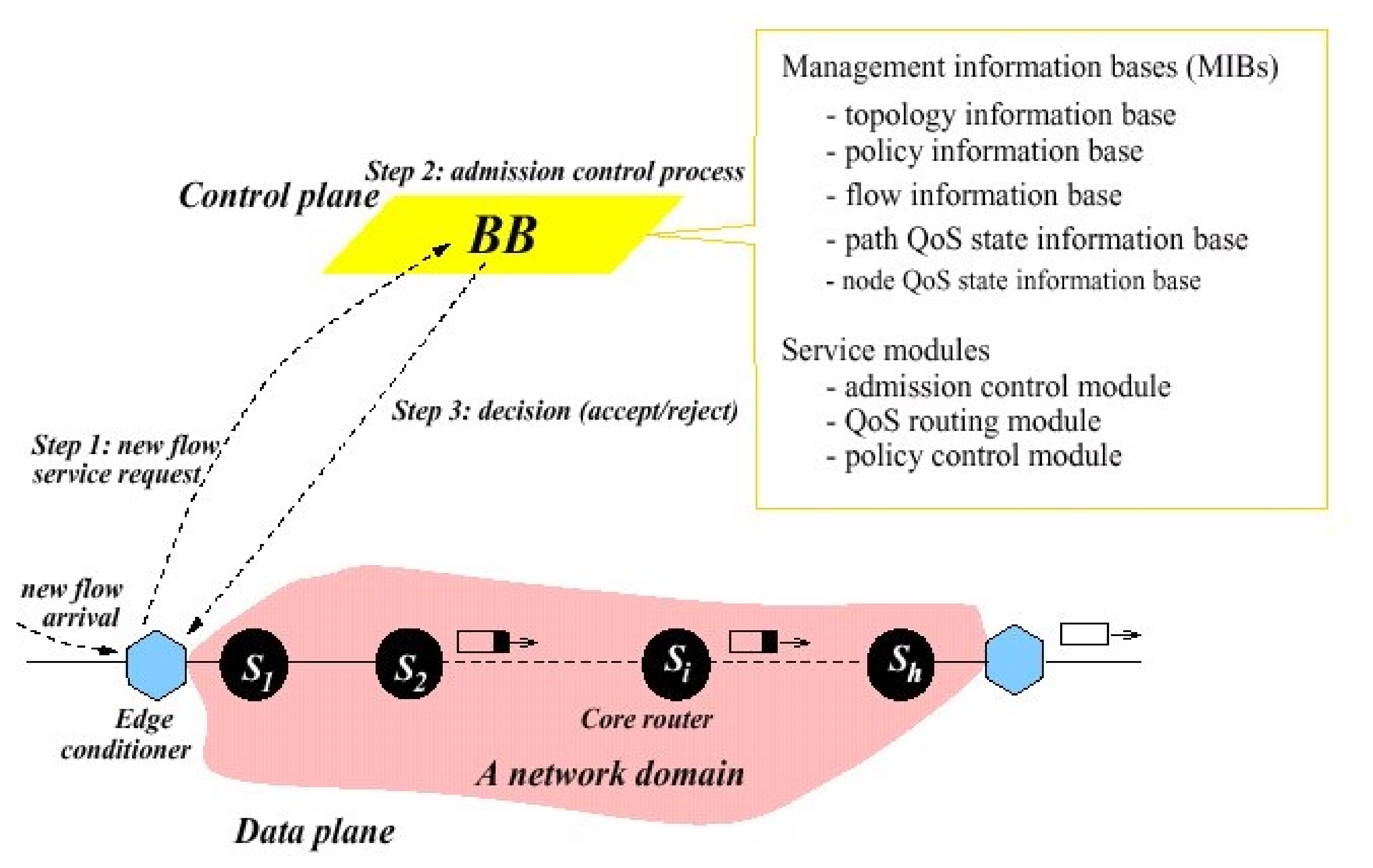
$$d_{edge}^j = \frac{\sigma^j - L^{j,\max}}{P^j - \rho^j}\frac{P^j - r^j}{r^j} + \frac{L^{j,\max}}{r^j} = T_{on}^j\frac{P^j - r^j}{r^j} + \frac{L^{j,\max}}{r^j}$$

end - to - end delay bound

$$\Rightarrow d_{e2e}^j = d_{edge}^j + d_{core}^j$$

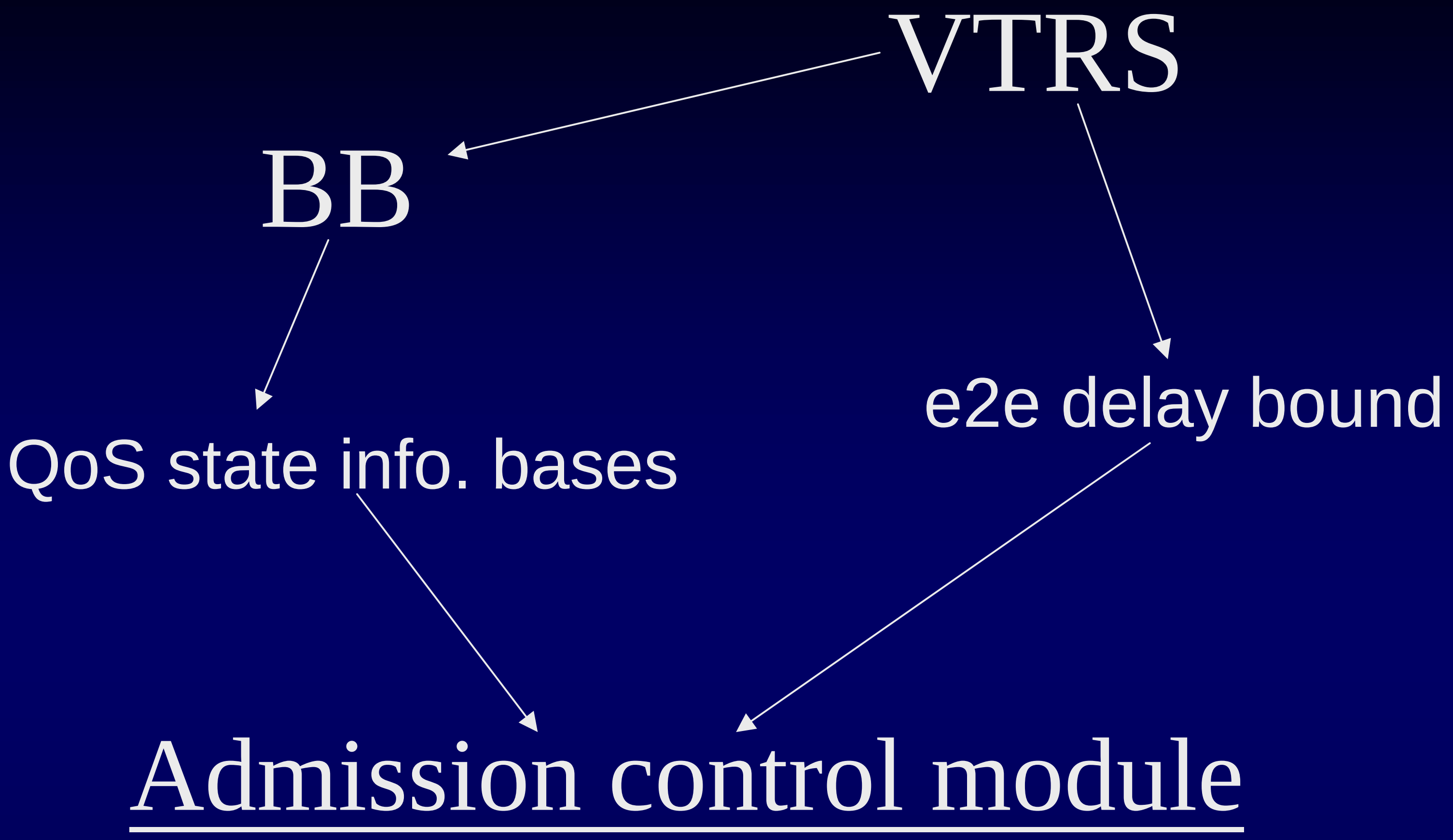
$$= T_{on}^j\frac{P^j - r^j}{r^j} + (q+1)\frac{L^{j,\max}}{r^j} + (h-q)d^j + \sum\nolimits_{i\in P}(\Psi_i + \pi_i)$$

Bandwidth Broker Architecture



QoS State Information Bases

- Flow information base
  - Flow id
  - Traffic profile: (σ^j, ρ^j, P^j, $L^{j,max}$)
  - Service profile: $D^{j,req}$
  - QoS reservation: (r,d)
- Node QoS state information base
  - Bandwidth, buffer capacity, scheduler type, error term
- Path QoS state information base
  - Hop number, sum of error terms and propagation delays, minimal residual bandwidth along the path



VTRS

BB

QoS state info. bases

e2e delay bound

Admission control module

Whether there is a feasible rate or not,
with which
delay requirement
is less than or equals to
e2e delay bound

Admission Control

- For per-flow guaranteed services
  - Pure rate-based schedulers
  - Mixed rate- and delay-based schedulers
  - Scalability?
- Dynamic flow aggregation
- For class-based guaranteed services
  - Pure rate-based schedulers
  - Mixed rate- and delay-based schedulers

Per-flow: Path with Only Rate-based Schedulers

- Parameters

$P: path$

$v: flow$

r^v : reserved rate of flow v

d^v : delay parameter of flow v

S_i : core router i

F_i : set of flows currently traversing S_i

C_i : total bandwidth of S_i

$C_{res}^{S_i}$: residual bandwidth at S_i

C_{res}^{P} : minimal residual bandwidth, i.e. $C_{res}^{P} = \min_{i \in P} C_{res}^{S_i}$

$(\sigma^v, \rho^v, P^v, L^{v,\max})$: traffic profile of flow v

$D^{v,req}$: end-to-end delay requirement

Per-flow: Path with Only Rate-based Schedulers

- Fundamental inequalities

$$\rho^{v} \le r^{v} \le P^{v} \quad \text{and} \quad r^{v} \le C^{P}_{res}$$

$$T^{v}_{on}\frac{P^{v}-r^{v}}{r^{v}} + (h+1)\frac{L^{v,\max}}{r^{v}} + \sum_{i\in P}\left(\Psi_i + \pi_i\right) \le D^{v,req}$$

- Feasible rate range derivation

Let $r^{v}_{\min}$ be the smallest r^{v}

$$\Rightarrow r^{v}_{\min} = \left[T^{v}_{on}P^{v} + (h+1)L^{v,\max}\right] / \left[D^{v,req} - \sum_{i\in P}\left(\Psi_i + \pi_i\right) + T^{v}_{on}\right]$$

Therefore, *feasible rate range*, R^{*}_{fea}, is defined as

$$\left[r^{low}_{fea}, r^{up}_{fea}\right] = \left[\max\left\{\rho^{v}, r^{v}_{\min}\right\}, \min\left\{P^{v}, C^{P}_{res}\right\}\right]$$

- The flow is admissible if the feasible rate range is non-empty, d^{v} is not necessary to be determined.
- The admissibility test can be done in O(1)

Per-flow: Path with Mixed Rate- and Delay-based Schedulers

- Fundamental inequalities

$$\rho^v \le r^v \le P^v,$$

$$r^v \le C_{res}^P,$$

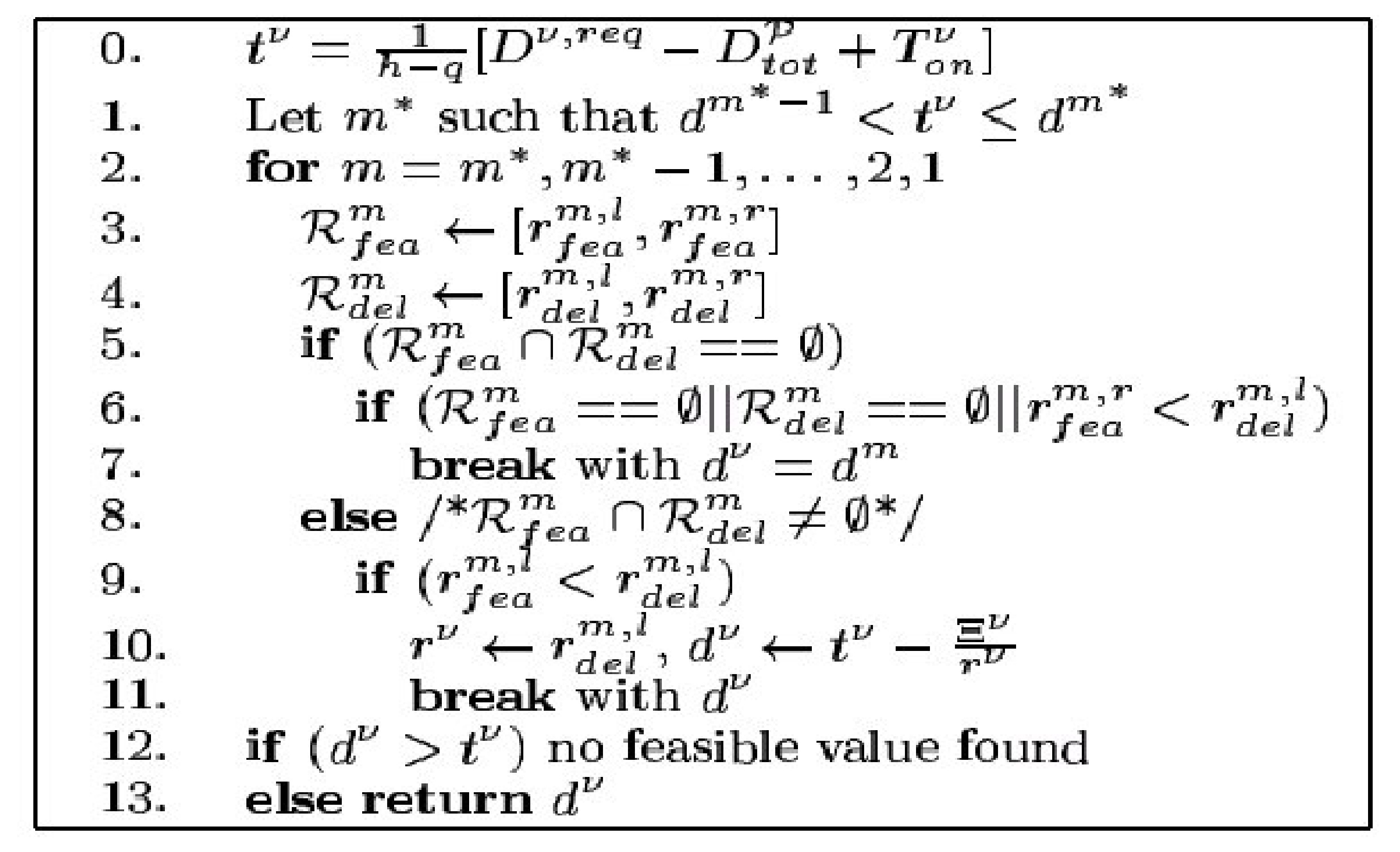
$$T_{on}^v \frac{P^v - r^v}{r^v} + (q+1)\frac{L^{v,\max}}{r^v} + (h-q)d^v + \sum_{i\in P}\left(\Psi_i + \pi_i\right) \le D^{v,req},$$

and

$$\sum_{\left\{\mathrm{j}\in \mathrm{F_i}:d_i^j \le d_i^k\right\}} \left[r^j\left(d_i^k - d_i^j\right) + L^{j,\max}\right] + \left[r^v\left(d_i^k - d^v\right) + L^{v,\max}\right] \le C_i d_i^k.$$

Per-flow: Path with Mixed Rate- and Delay-based Schedulers

- Efficient algorithm :



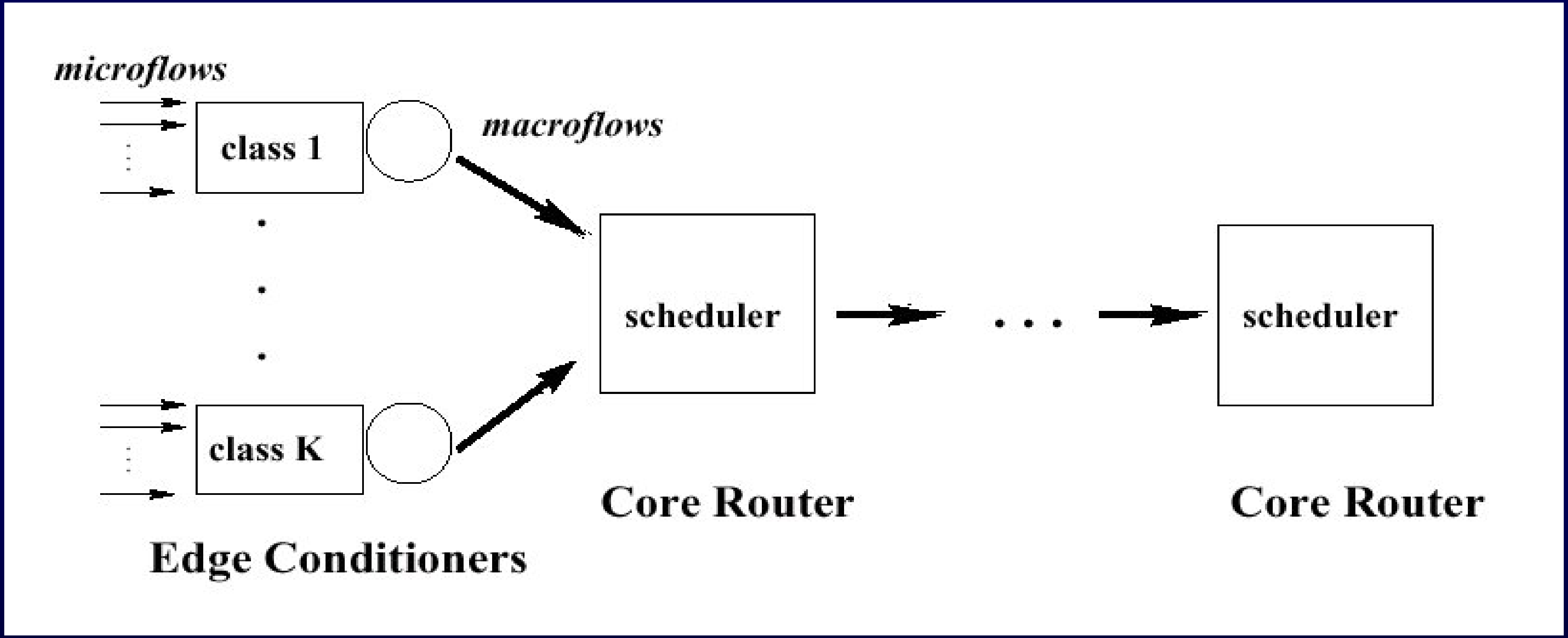
```
$t^{\nu} = \frac{1}{h-q}[D^{\nu,req} - D^{p}_{tot} + T^{\nu}_{on}]$
Let $m^*$ such that $d^{m^*-1} < t^{\nu} \le d^{m^*}$
for $m = m^*, m^*-1, \ldots, 2, 1$
    $\mathcal{R}^{m}_{fea} \leftarrow [r^{m,l}_{fea}, r^{m,r}_{fea}]$
    $\mathcal{R}^{m}_{del} \leftarrow [r^{m,l}_{del}, r^{m,r}_{del}]$
    if $(\mathcal{R}^{m}_{fea} \cap \mathcal{R}^{m}_{del} == \emptyset)$
        if $(\mathcal{R}^{m}_{fea} == \emptyset || \mathcal{R}^{m}_{del} == \emptyset || r^{m,r}_{fea} < r^{m,l}_{del})$
            break with $d^{\nu} = d^{m}$
    else $/* \mathcal{R}^{m}_{fea} \cap \mathcal{R}^{m}_{del} \neq \emptyset */$
        if $(r^{m,l}_{fea} < r^{m,l}_{del})$
            $r^{\nu} \leftarrow r^{m,l}_{del}$, $d^{\nu} \leftarrow t^{\nu} - \frac{\Xi^{\nu}}{r^{\nu}}$
            break with $d^{\nu}$
if $(d^{\nu} > t^{\nu})$ no feasible value found
else return $d^{\nu}$
```

- Time complexity

$$O(M), \text{ where } M \le \sum_{S_i \text{ id delay-based}} |F_i|.$$

Class-based Guaranteed Service Model

- Enhance the scalability of proposed BB architecture
- Service Model



- Dynamic flow aggregation has not been identified nor addressed

Dynamic Flow Aggregation (1/5)

- Impact on e2e delay (macroflow α → α')
  - All rate-based schedulers

worse - case delay at edge conditioner is larger than

$$d_{edge}^{\alpha'} = T_{on}^{\alpha'} \frac{P^{\alpha'} - r^{\alpha'}}{r^{\alpha'}} + \frac{L^{\alpha',\max}}{r^{\alpha'}}$$

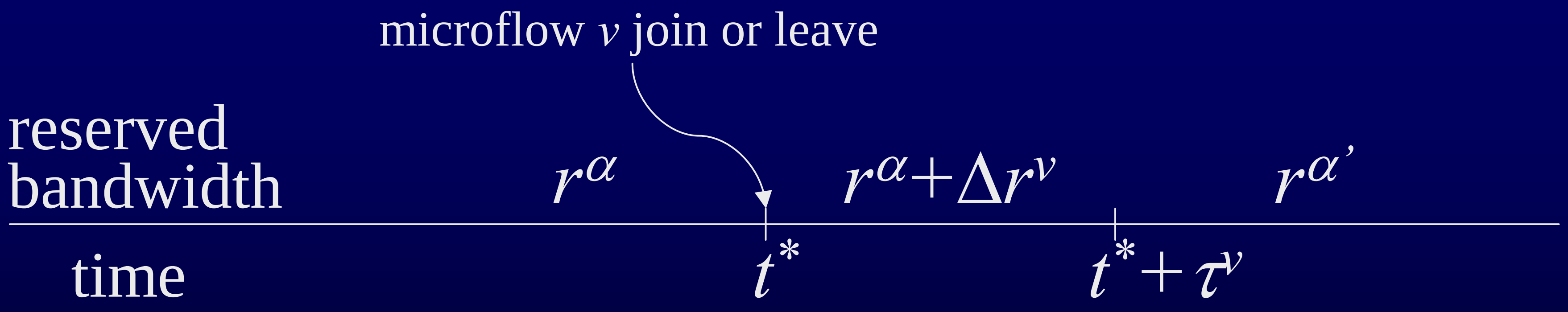
some packets from new macroflow may experience a worst - case delay in the network core

by $d_{core}^{\alpha} = h \frac{L^{P,\max}}{r^{\alpha}} + \sum_{i \in P} \left(\Psi_i + \pi_i \right)$

instead of $d_{core}^{\alpha'} = h \frac{L^{P,\max}}{r^{\alpha'}} + \sum_{i \in P} \left(\Psi_i + \pi_i \right)$

Dynamic Flow Aggregation (2/5)

- Edge delay bound
  - Contingency bandwidth : to eliminate the lingering delay effect of the backlog packets
  - A new microflow v aggregates or de-aggregates, the contingency bandwidth is Δr^{v}, and the contingency period is τ^{v}.



microflow v join or leave

reserved bandwidth

r^{α}

$r^{\alpha}+\Delta r^{v}$

$r^{\alpha'}$

time

t^{*}

$t^{*}+\tau^{v}$

  - Δr^{v} and τ^{v} are chosen to bound the edge delay as

$$d_{edge}^{new} \le \max\left\{d_{edge}^{\alpha}, d_{edge}^{\alpha'}\right\}$$

Dynamic Flow Aggregation (3/5)

- Edge delay bound
  - The microflow v is with ($\sigma^v, \rho^v, P^v, L^{v,max}$).
  - Sufficient conditions on Δr^v and τ^v :

$$\begin{cases} \Delta r^v \geq P^v - r^v & \text{(microflow join)} \\ \Delta r^v \geq r^v & \text{(microflow leave)} \end{cases}$$

and $\tau^v \geq \dfrac{Q(t^*)}{\Delta r^v}$,

where $Q(t^*) \leq d_{edge}^{\alpha}\left(r^{\alpha} + \Delta r^{\alpha}(t^*)\right)$ is the backlog,

where $\Delta r^{\alpha}(t^*)$ is the total contingency bandwidth allocated to the macroflow α at time t^*.

→ *contingency period bounding*

Dynamic Flow Aggregation (4/5)

- Core delay bound

$$d_{core}^{\alpha'} = q\max\left\{\frac{L^{P,\max}}{r^{\alpha}},\frac{L^{P,\max}}{r^{\alpha'}}\right\} + (h-q)d^{\alpha} + \sum\nolimits_{i\in P}\left(\Psi_i + \pi_i\right)$$

Dynamic Flow Aggregation (5/5)

- Admission control: Microflow join

$$d^{\alpha'}_{e2e} = d^{\alpha'}_{edge} + \max\left\{ d^{\alpha}_{core}, d^{\alpha'}_{core} \right\} \le D^{\alpha,req}$$

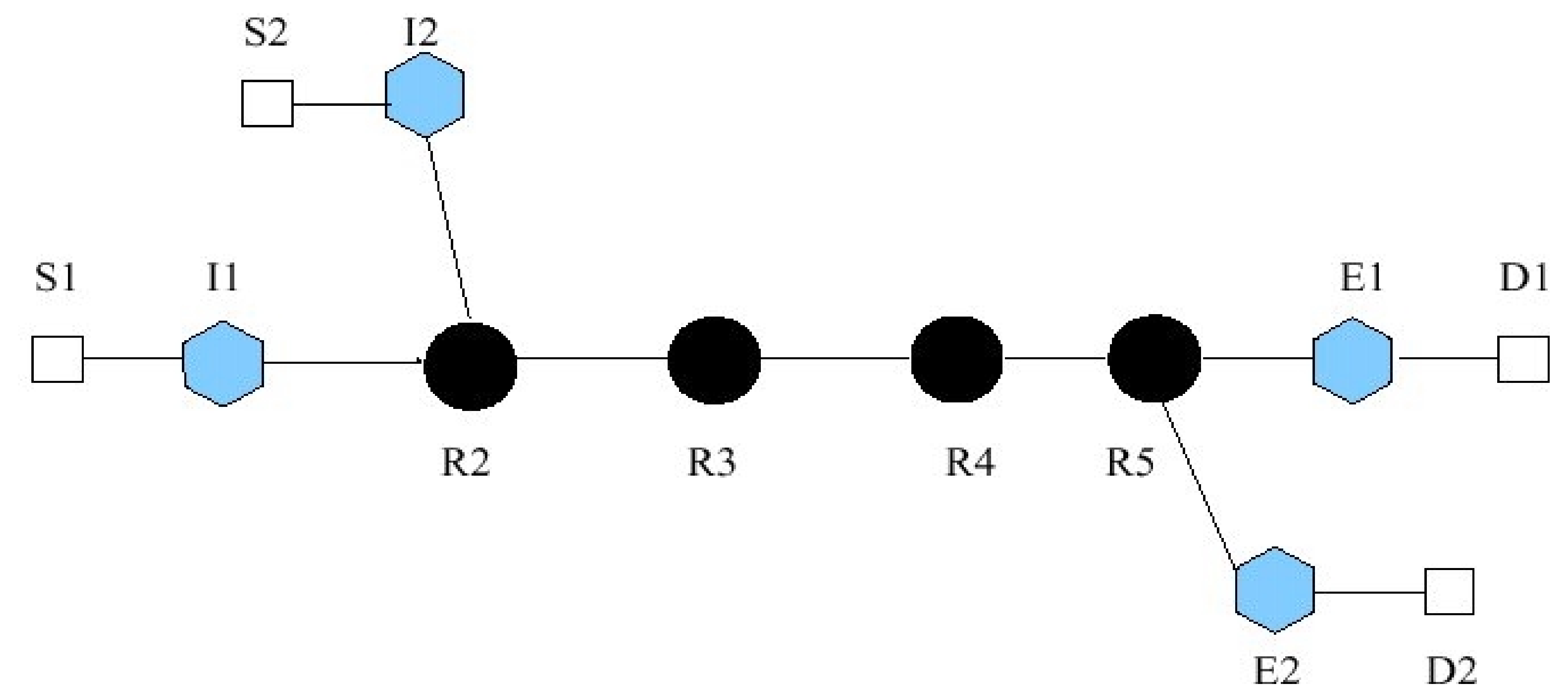
since $r^{\alpha'} \ge r^{\alpha}$, hence, $d^{\alpha}_{core} \le d^{\alpha'}_{core}$.

$\Rightarrow d^{\alpha'}_{edge} \le D^{\alpha,req} - d^{\alpha}_{core}$.........(a)

also, $\rho^{\mathrm{v}} \le r^{\alpha'} - r^{\alpha} \le P^{v}$........(b)

$\Rightarrow r^{\alpha'}$ can be derived from (a) (b)

Simulation Investigation



Comparison

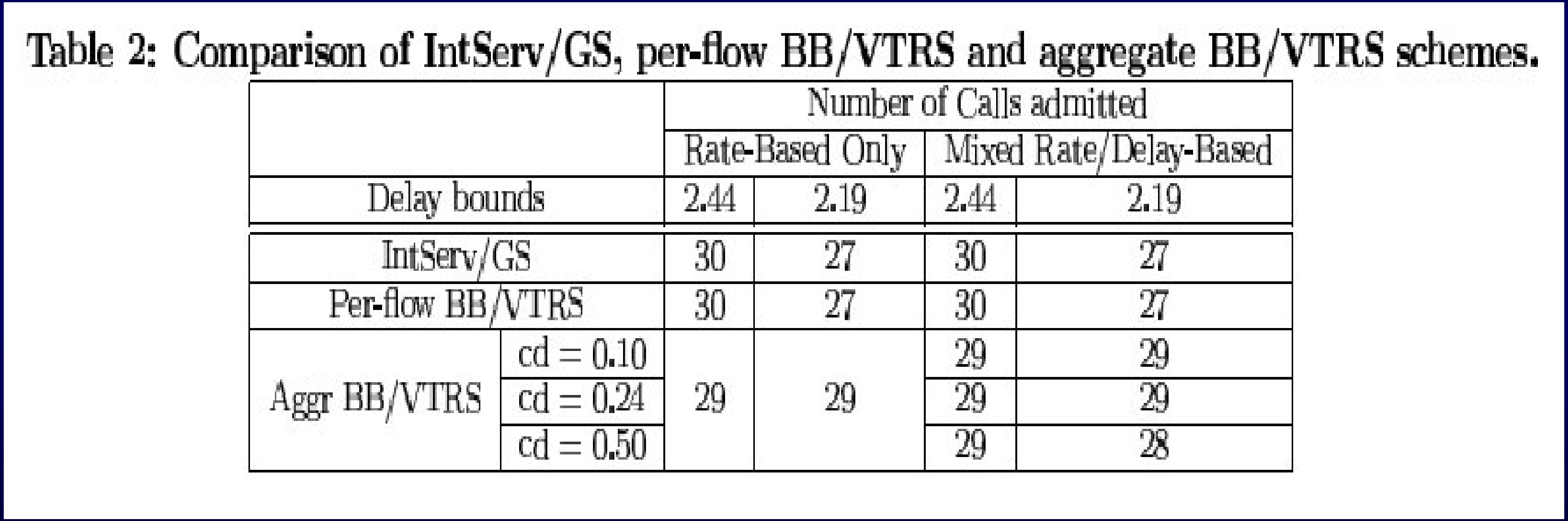
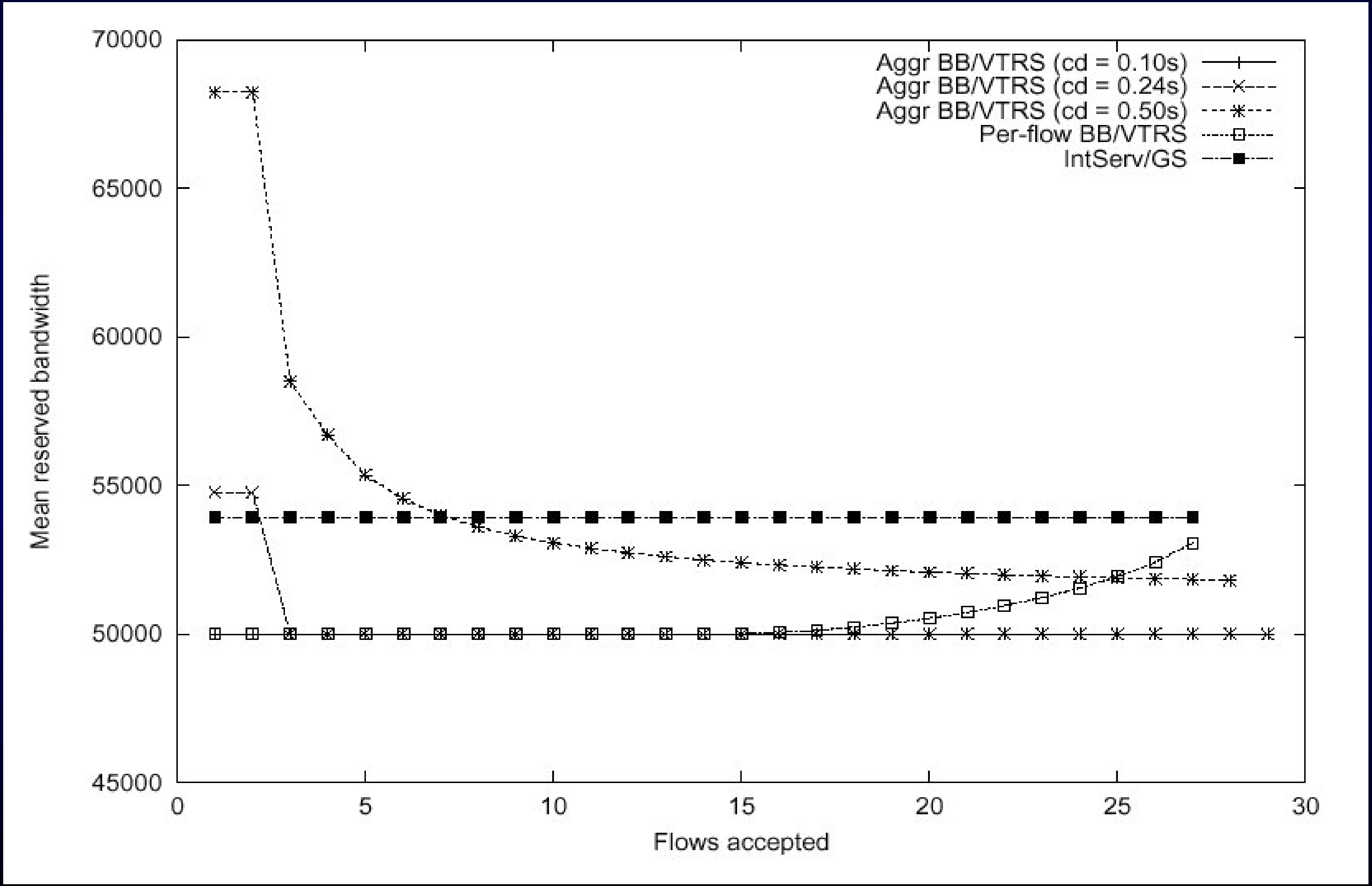


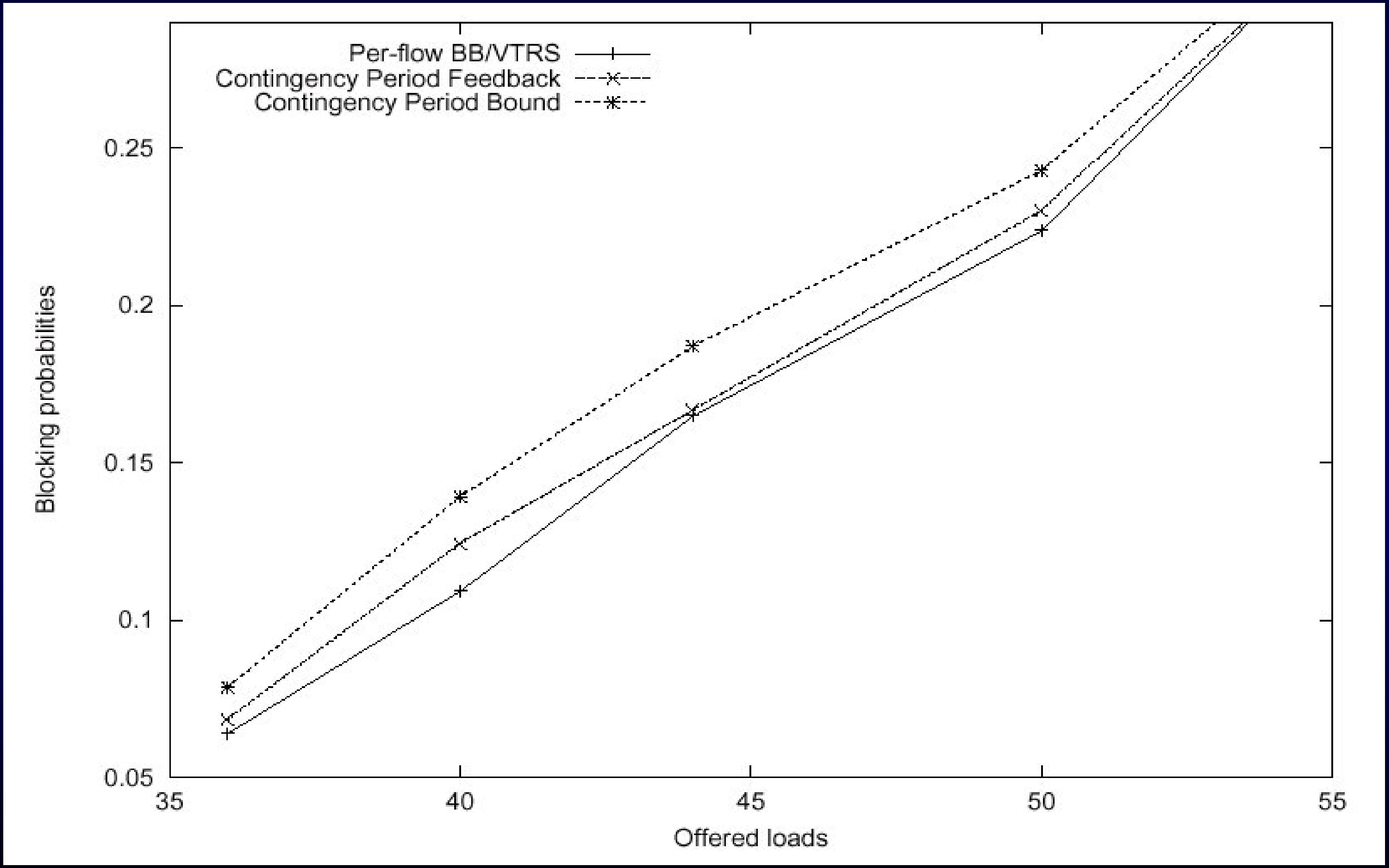
Table 2: Comparison of IntServ/GS, per-flow BB/VTRS and aggregate BB/VTRS schemes.

		Number of Calls admitted			
		Rate-Based Only		Mixed Rate/Delay-Based	
Delay bounds		2.44	2.19	2.44	2.19
IntServ/GS		30	27	30	27
Per-flow BB/VTRS		30	27	30	27
Aggr BB/VTRS	cd = 0.10	29	29	29	29
	cd = 0.24			29	29
	cd = 0.50			29	28

Mean Reserved Bandwidth



Flow Blocking Rate



Conclusion

- Present a novel BB architecture based on VTRS
- Decouple the QoS control plane from data plane
- Propose path-oriented admission control approach
- Support per-flow and class-based guaranteed services
- No or minimal configuration of core routers

Future Works

- Distributed bandwidth broker architecture
- Inter-Domain QoS reservation and service level agreement