Fair Scheduling Scheme with Feedback in the Joint Allocation of Heterogeneous Resources

Yu Hua
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan, China, 430074
Email: yhuastar@hotmail.com

***Abstract*— Efficient QoS guarantee in network environments mainly relies on the fair scheduling and reliable allocation of computing, storage space and bandwidth resources, which are available in each node. However, most works focus on the management of bandwidth allocation and special resources, such as wavelengths in optical networks, channels allocation in wireless networks and energy management in wireless sensor networks. It is very important that all kinds of resources in each network node are considered together according to the current network state.**

In this paper, a kind of novel architecture is presented, which supports the joint allocation of heterogeneous resources. The practical algorithm is called Deficit Round Robin with Feedback Constraints (DRR-FC) and its basic idea is to model the feedback information as constraints and schedule heterogeneous resources in the DRR-based way. The whole process emphasizes the cooperative relationships among different resources. Based on the network calculus theory, the deterministic performance bounds are presented and through the simulations, the novel architecture is proved efficient and feasible in current network environments.

***Keywords*-Fair Scheduling, Joint Allocation, Network Calculus, Feedback Constraints, Deficit Round Robin**

I. INTRODUCTION

A. Background

Generally, network traffics are rather unstable and hard to predict in the dynamic network environments. Flows with different attributes can put forward different QoS requirements, which are described by all kinds of QoS parameters. The attributes of flows may be latency-critical, throughput-critical and so on. Meanwhile, the end-to-end transmissions need not only the single resource (*i.e.* the bandwidth), but also all kinds of heterogeneous resources shared, such as the processor, buffer space, capacity of power, and even the large software used for network management. Hence, the efficient policy needs to pay more attentions on all kinds of heterogeneous resources, not a single resource.

Current scheduling algorithms can be classified as sorted-priority and frame-based methods. Sorted-priority algorithms transmit packets in increasing order according to the time-stamp, which is defined in each queued packet. The typical sorted-priority algorithm is the Weighted Fair Queuing (WFQ) and the corresponding practical methods include Packet Generalized Processor Sharing (PGPS) [3], Self-Clocked Fair Queuing (SCFQ) [4], Worst-case Fair weighted Fair Queuing (WF2Q) [5], and the schemes based on the Rate Proportional Server (RPS) framework [6], [7]. Among them, WF2Q has the best performance, but it is hard to be implemented due to the high complexity especially in the high-speed networks. Furthermore, WF2Q+ algorithm was proposed in [8]. The complexity of WF2Q+ was lower because of the reduction on updating the virtual clocks.

In addition, in the frame-based algorithms, time splice is divided into several frames and data packets can be transmitted with the per-frame format. Deficit Round Robin (DRR) [9] is a credit-based extension of classical Weighted Round Robin (WRR) with lower complexity than WFQ. DRR has O(1) worst-case per packet complexity if the operation number is constant with respect to the number of active flows [10]. Corresponding improvements include other round-robin algorithms such as Pre-order Deficit Round-Robin (PDRR) [11], Heterogeneous-Deficit Round Robin (H-DRR) [12] and Smoothed Round-Robin (SRR) [18]. As stated in [9] and [10], the flows in DRR can share the bandwidth fairly with variable packet lengths.

B. Related Work

The ideas about fairness mainly focus on the allocation of single resource (such as bandwidth) on a link or multiple resources of the same kind. As mentioned above, arriving flows need to utilize different kinds of resources, such as processor, bandwidth, and buffer space, all together. The essence of end-to-end QoS guarantee is based on the fair allocation of heterogeneous resources in order to improve the resources utilization, increase the throughput and reduce transmission delay.

Taking into account the characteristics of channel fading and intercell interference, Liang Xu, Xuemin Shen, Mark, J.W. [13] used the generalized processor sharing (GPS) fair service discipline to allocate uplink channel-resources in order to efficiently support real-time and non-real-time multimedia traffic with guaranteed statistical QoS in the uplink of a wideband code-division multiple access (CDMA) cellular network. In unstructured peer-to-peer systems, Drougas. Y.,

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

Kalogeraki. V. [14] utilized the fairness index of a distribution as a measure of fairness and showed the methods to optimize the fairness of the distribution using only local decisions. However, the works above mainly focus on the fair allocation of the single resource in the special network environments.

Yunkai Zhou and Harish Sethu [1],[2] investigated the issue of achieving fairness in the joint allocation of the processing and bandwidth resources and provided a shared processor and a shared link with max-min fairness. Their works are important to realize the fair and joint allocation of the heterogeneous resources. The authors appended other kinds of resources as new parts to existing scheduling algorithms, *i.e.*, Deficit Round Robin.

C. Contributions

In the current environments, network tasks need all kinds of resources working cooperatively, not a single resource. Meanwhile, the bottleneck resource changes with the uncertain network state. Hence, the allocation of the heterogeneous resources should be fair and cooperative in order to provide the efficient scheduling policy timely.

The contributions in this paper can be listed as follows. First, a novel architecture is presented, which can support the fair and joint allocation of the heterogeneous resources. The architecture is different from the previous works and it utilizes the feedback information of corresponding queues to cooperatively adjust the allocation policy. Second, the DRR-FC algorithm is provided in order to achieve the fair allocation and in this algorithm, every kind of resource is considered as a adjustable factor, not a static entity like [2]. The improvement can obtain the more flexible allocation policy. Third, based on the network calculus theory, the deterministic bounds are analyzed and the performance of the novel architecture can be guaranteed in theory.

D. Organization

The paper is organized as follows. In section II, a novel architecture is proposed, which can support the joint allocation of heterogeneous resources. The scheduling algorithm is presented in section III. In section IV, the mathematical analysis based on the network calculus theory is shown, which gives the performance bounds of proposed methods. The results and analysis of the simulations can be presented in section V. Finally, the conclusion is given in section VI.

II. Proposed Architecture and Mechanism

A. Fairness Measure

Generalized Processor Sharing (GPS) [5] theory is the classical model for achieving the mechanism and methods of fair scheduling. GPS server can be defined as:

$$\frac{w_i(t_1,t_2)}{w_j(t_1,t_2)} \geq \frac{\theta_i}{\theta_j}, i,j = 1,\ldots,N$$

In this model, $w_i(t_1,t_2)$ and $w_j(t_1,t_2)$ are the transmitted amounts for flow i and j during the time interval $[t_1,t_2]$. θ_i and θ_j are the positive real numbers and stand for the flow i and j, which are served by GPS servers. In order to execute fair allocation based on GPS model, arriving packets need to be separated into infinitesimal parts.

The fair measure based on the feedback information is given in order to support the fair allocation of multiple heterogeneous resources. As for the data aggregation x, the model of fairness measure can be defined as:

$$\max\left\{\left|\frac{c_x(t_1,t_2)}{\omega_c(x)} - \frac{s_x(t_1,t_2)}{\omega_s(x)}\right|, \left|\frac{s_x(t_1,t_2)}{\omega_s(x)} - \frac{b_x(t_1,t_2)}{\omega_b(x)}\right|\right\} < \varepsilon$$

In this model, $c_x(t_1,t_2)$, $s_x(t_1,t_2)$ and $b_x(t_1,t_2)$ stand for the service provided by computation, storage and bandwidth resources during the time interval $[t_1,t_2]$, respectively. In addition, $\omega_c(x)$,$\omega_s(x)$ and $\omega_b(x)$ for the data aggregation x are the adjustable weights based on the feedback information for computation, storage and bandwidth resources. ε is a infinitesimal number. Thus, for the single resource, the ratio of the service provided and the weights assigned can be considered as the fairness index of the current resource. Then, the index differentiation among different resources can be computed and the maximum is used to evaluate the fairness.

B. Proposed Architecture

In each network node, there are all kinds of heterogeneous resources available, such as wavelengths of optical nodes, the channels and energy of wireless sensor node. Here, the proposed architecture focuses on the capacities of the bandwidth, storage and computation. The architecture is scalable and can support other resources as well.

The architecture is shown in Fig.1, which can support the fair and joint scheduling of computation, storage and bandwidth resources according to the feedback information. In essence, the architecture is composed of three parts: information collection, feedback constraints processing and joint scheduling.

The state information of heterogeneous resources can be collected from corresponding queues and sent to feedback constraints producer. The state information includes the length of the waiting queue and can show whether the current resources are busy. Afterwards, according to the current state information, the feedback constraints producer adjusts the policy of allocating resources based on the service level agreement and transmits the constraints to the converter. In the constraints converter, the constraints are classified into different resources types and sent to respective resources managers, such as the rate controller, bandwidth and storage allocation. Thus, each resource manager can set new allocation policy according to the arriving constraints and all the execution solutions from different resources managers can be gathered in the

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

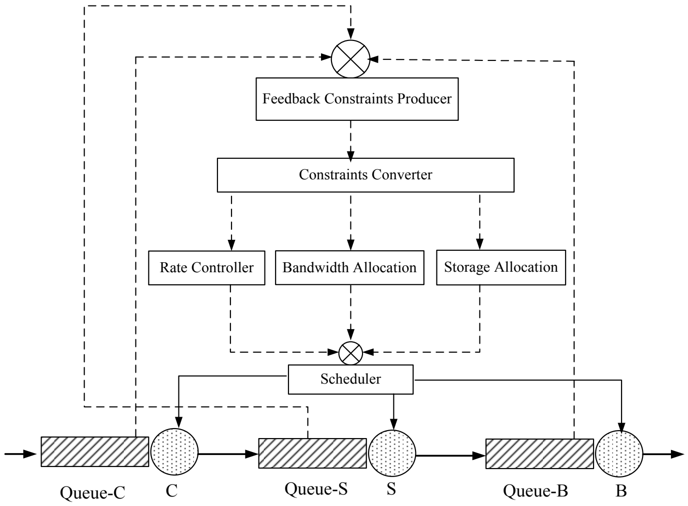

Fig. 1. The proposed architecture supporting the fair and joint allocation of heterogeneous resources. C, S and B stand for the resource capacities of Computation, Storage and Bandwidth, respectively.

scheduler, which is responsible for performing the cooperative and unified allocation of the heterogeneous resources.

III. Scheduling Algorithm

In this part, the Deficit Round Robin with Feedback Constraints (DRR-FC) algorithm is presented in order to support the proposed architecture. The basic idea of DRR-FC algorithm is to allocate and schedule the heterogeneous resources one by one based on the allocated feedback weights in the deficit round robin way. The DRR-FC algorithm consists of three parts: *Initialization, Inqueue and Dequeue* models.

The algorithm mainly focuses on three kinds of resources, *computation, storage (buffer space) and bandwidth,* in each network node. The practical aim is to make the arriving data join different queues, in which the resources can be fairly allocated with the feedback constraints in the deficit round robin way.

The Initialization and Inqueue models are shown in Fig.2. In the Initialization model, the lists are set to *null* in order to temporarily store the waiting data and three kinds of the quanta, *i.e.*, DC_c, DC_s and DC_b, are set to zero for allocating the feedback-based quantum in each round.

As for the data aggregation x, the *enqueue* model is responsible for making the requirements of different resources capacities join corresponding waiting queues. Of course, the length check should be done in order to avoid the overflow in each queue. When the requirements of the current data aggregation are in waiting queues, corresponding quanta need to be set zero.

The main part of DRR-FC algorithm is the *Dequeue* model presented in Fig.3 and its function is to fairly schedule the

DRR-FC Algorithm $\{$
Initialization Model
$List_c, List_s, List_b \longleftarrow null$
$DC_c, DC_s, DC_b \longleftarrow 0$
Enqueue Model
Get the arriving data aggregation $\boldsymbol{x}$
if $Length(x) > Threshold$ **then**
  reject x
end if
if $ExistinLists(x) == False$ **then**
  add x into $List_c(x_i), List_s(x_j), List_b(x_k)$
  $DC_c(x_i), DC_s(x_j), DC_b(x_k) \longleftarrow 0$
end if

Fig. 2. The initialization and enqueue models of Deficit Round Robin with Feedback Constraints algorithm. ω_c, ω_s and ω_b are the feedback weights.

resources available in order to satisfy the requirements in each queue. As for the computation resource, it has a waiting queue, which contains all kinds of requirements for the computation resources. The element in the head of the list is checked whether its DC_c value is larger than the current processing cost, after the increment by the feedback weight, ω_c. If it is true, the system will schedule resources, execute related operations and subtract the consumed processing cost from the deficit counter. If it is false, the quantum will be added into the next round and the element is inserted into the list again. The same operations are applied to the other two kinds of resources, storage and bandwidth, as well. When the current requirements of data aggregation x have been satisfied after the operations in three lists, x can be added into respective lists again if there are other requirements.

IV. Mathematical Analysis

Network calculus has been considered as a useful method to provide deep insights into network traffic shaping and scheduling in nodes. Network calculus is a kind of novel system theory, which is based on Min-Plus algebra and can be used in computer networks [15]. The underlying idea is that service guarantees can be achieved by regulating the traffic and deterministic scheduling [16]. In this part, the basic performance bounds of the proposed architecture are presented with the network calculus theory.

A. Definition

definition 1 (Service Curve): Consider a system S and a flow through S with input and output function F and $\tilde{F}$. $\xi(t)$ is a nonnegative wide-sense increasing function. The service curve is ξ if and only if $\tilde{F}(t) - F(\tau) \geq \xi(t-\tau)$ with certain $0 \leq \tau \leq t$. Based on the min-plus principle,

$$\tilde{F}(t) \geq \inf_{0<\tau\leq t}\{\xi(t-\tau) + F(\tau)\} = (\xi \otimes F)(t)$$

Furthermore, according to the definition of the service curve, the improved service curve can be given, which is based on the feedback constraints.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

Dequeue Model
{
while (***true***) **do**
  if $List_c \neq null$ **then**
    $x_i \longleftarrow HeadofList_c$
    Remove x_i from $List_c$
    $DC_c(x_i) := DC_c(x_i) + \omega_c(x_i)$
    if $ProcessCost(x_i) > DC_c(x_i)$ **then**
      Insert x_i into the tail of $List_c$ and break
    end if
    Schedule computation resources for x_i
    $DC_c(x_i) := DC_c(x_i) - ProcessCost(x_i)$
  end if
  if $List_s \neq null$ **then**
    $x_j \longleftarrow HeadofList_s$
    Remove x_j from $List_s$
    $DC_s(x_j) := DC_s(x_j) + \omega_s(x_j)$
    if $StorageCost(x_j) > DC_s(x_j)$ **then**
      Insert x_j into the tail of $List_s$ and break
    end if
    Schedule storage resources for x_j
    $DC_s(x_j) := DC_s(x_j) - StorageCost(x_j)$
  end if
  if $List_b \neq null$ **then**
    $x_k \longleftarrow HeadofList_b$
    Remove x_k from $List_b$
    $DC_b(x_k) := DC_b(x_k) + \omega_b(x_k)$
    if $BandwidthCost(x_k) > DC_b(x_k)$ **then**
      Insert x_k into the tail of $List_b$ and break
    end if
    Schedule bandwidth resources for x_k
    $DC_b(x_k) := DC_b(x_k) - BandwidthCost(x_k)$
  end if
  if $EmptyofRequire(x)$ and $OtherRequire(x)$ **then**
    insert x into $List_c(x_i), List_s(x_j), List_b(x_k)$
  end if
end while

Fig. 3. The dequeue model of the Deficit Round Robin with Feedback Constraints algorithm. ω_c, ω_s and ω_b are the feedback weights from the scheduler.

definition 2 (Service Curve with Feedback Information): The coefficient of service rate R for computation resource is

$$\eta_i = \frac{\omega_c(x_i)}{\frac{\sum_{i=1}^{N} \omega_c(x_i)}{N}}$$

The coefficient of the system delay T for computation resource can be described as

$$\lambda_i = 1 - \frac{\omega_c x_i}{\sum_{i=1}^{N} \omega_c x_i}$$

Based on the basic definition, the service curve ϑ_c for computation resources of data x_i can be defined as

$$\vartheta_c = \eta_i R + \lambda_i T$$

The service curves of other resources, such as storage and bandwidth, can be defined in the similar way.

B. Lemma

lemma 1 (System Delay): For data x_i, the whole delay D is smaller than the sum of the delay for computation resource, D_c, and the delay for bandwidth, D_b.

Proof:

Consider the VBR flows and the *affine* arrival curve:

$$\theta_{r,b}(t) = rt + b$$

Based on the definition 2,

$$D = \frac{b}{(\eta_i R_c) \wedge (\eta_k R_b)} + \lambda_i T_c + \lambda_k T_b$$

Meanwhile, as for scheduling computation resource, the output is constrained by

$$\theta_c(t) = b + rt + \lambda_i r T_c$$

Furthermore,

$$D_c = \frac{b}{\eta_i R_c} + \lambda_i T_c$$

In the same way, get

$$D_b = \frac{b + \lambda_i T_c}{\eta_k R_b} + \lambda_k T_b$$

Therefore, it is obvious that

$$D < D_c + D_b$$

lemma 2 (Queue Size): In the fair scheduling architecture with feedback constraints, the size of waiting queue for computation resource can be bounded by

$$\max\{\frac{rT_c + b}{T_c}, \frac{\sum_{i=1}^{N} \omega_c(x_i)}{N}\}T_c + b$$

Proof:

Consider the *rate-latency* service curve,

$$\delta_{R,T}(t) = R[t-T]^+ = \begin{cases} R(t-T), & \text{if } t > T \\ 0, & \text{otherwise} \end{cases}$$

Hence, in the worst case, the arriving rate relies on the comparison between the arriving data and average capacity in each time splice. Then, the rate can be bounded by

$$\max\{\frac{rT_c + b}{T_c}, \frac{\sum_{i=1}^{N} \omega_c(x_i)}{N}\}$$

Afterwards, the maximum of the queue size is

$$\max\{\frac{rT_c + b}{T_c}, \frac{\sum_{i=1}^{N} \omega_c(x_i)}{N}\}T_c + b$$

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

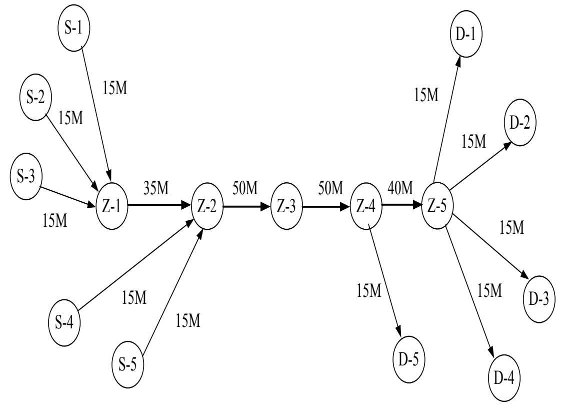

Fig. 4. The network topology for verifying DRR-FC performance. The number of each link stands for the basic bandwidth assigned.

V. Performance Study

A. Network Topology

The efficiency of the proposed DRR-FC algorithm can be verified through the simulations based on NS-2 [17]. The network topology is shown in Fig.4.

The source nodes are five nodes, in which the traffic can be generated according to ON (15 or 50 packets/s) and OFF (no packets) periods independently and identically distribution. Seven sessions are established between the source and destination nodes and can be denoted by source-destination pairs in the following table.

TABLE I
THE SESSION DESCRIPTIONS AMONG THE NETWORK NODES

	$Z_1 - Z_2$	$Z_2 - Z_3$	$Z_3 - Z_4$	$Z_4 - Z_5$
$S_1 - D_3$	22	22	22	22
$S_4 - D_5$	-	15	15	-
$S_2 - D_1$	34	34	34	34
$S_3 - D_4$	27	27	27	27
$S_1 - D_5$	18	18	18	-
$S_4 - D_3$	-	30	30	30
$S_5 - D_4$	-	41	41	41

B. Transmission Delay

The transmission delay is an important index for evaluating the performance of scheduling algorithm. When the traffic generator sends 15 packets/s, the whole description of the transmission delay is illustrated in Fig.5. The whole process lasts for 400 ms and four scheduling algorithms, DRR-FC, PPLS, WF2Q and WRR, are compared with the same traffic. The average delay are 0.342ms, 0.196ms, 0.117ms and 0.035ms, respectively. From the results above, it is proved that the joint scheduling solutions of DRR-FC and PPLS can be more efficient than the single solution of WRR and WF2Q. The joint allocation of heterogeneous resources is also helpful to avoiding the transmission congestion due to considering the computation, storage space and bandwidth resources together. Meanwhile, although both DRR-FC and PPLS belong to the joint scheduling policy, DRR-FC has the smaller delay than PPLS. DRR-FC can obtain the better transmission performance with the smaller delay. The main reason is that DRR-FC can utilize the feedback constraints to adapt the scheduling policy, which is more efficient to the burst traffic.

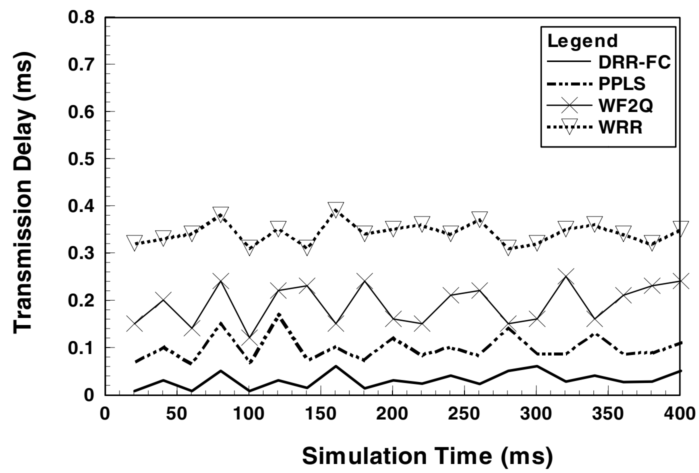

Fig. 5. The comparisons among different scheduling schemes, DRR-FC, PPLS, WF2Q and WRR, for the delay. The traffic model is according to ON (15 packets/s) and OFF (no packets) periods.

When the burst traffic is incremented to 50 packets/s, the comparisons are shown in Fig.6. Generally, the increased traffic leads to the larger delay, which is also illustrated in this figure. However, because the feedback constraints are considered and the adaptation is executed, the effect of increased traffic to DRR-FC is rather limited and little fluctuation can be produced in the whole simulation process.

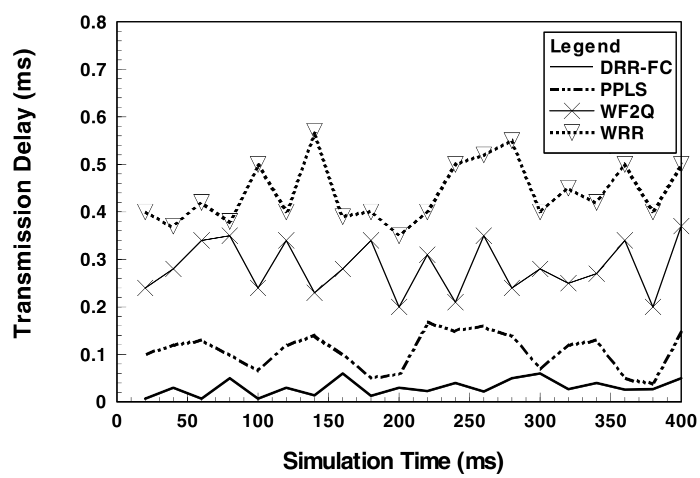

Fig. 6. The comparisons among different scheduling schemes, DRR-FC, PPLS, WF2Q and WRR, for the delay with 50 packets/s in ON state

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

C. *Fairness Measure*

One of the aims of scheduling algorithms is to realize the fair scheduling scheme. In order to check the fair efficiency of the proposed algorithm, the fair measure can be evaluated in the simulation. *Fairness Gap* refers to the absolute deviation between the actual and standard fairness measure as mentioned above. The fairness gaps of four scheduling algorithms can be illustrated in Fig.7 and 8, which present the results under different traffic models, 15 and 50 packets/s in ON state.

Under different traffic models, DRR-FC can achieve the smallest fairness gap in the whole process. The average values of fairness gap in Fig.7 are 1.26%, 5.74%, 13.57% and 27.9% for DRR-FC, PPLS, WF2Q and WRR, respectively. The similar trends can also be obtained in Fig.8. Therefore, DRR-FC can provide the better fairness for the joint allocation of heterogeneous resources.

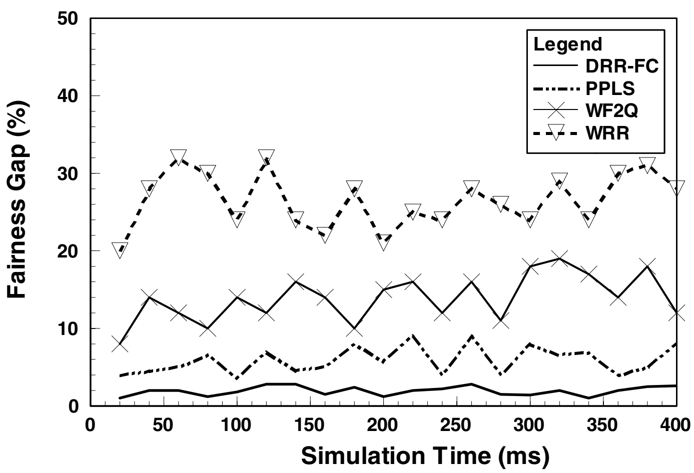

Fig. 7. The fairness gaps of DRR-FC, PPLS, WF2Q and WRR, with 15 packets/s in ON state.

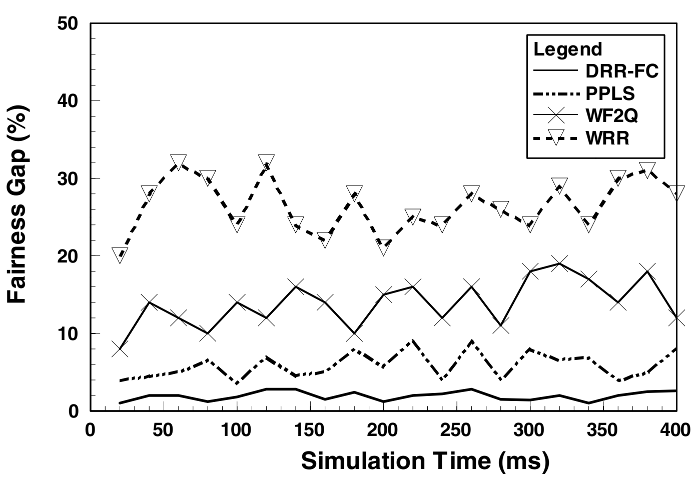

Fig. 8. The fairness gaps of DRR-FC, PPLS, WF2Q and WRR, with 50 packets/s in ON state.

VI. Conclusion

In this paper, a novel architecture is proposed, which can realize the fair allocation of heterogeneous resources. Meanwhile, the practical algorithm, DRR-FC, is also presented in order to support the efficient resources management based on the feedback constraints. The basic idea is to achieve the fair scheduling by jointly considering the feedback information of resources state. Furthermore, the mathematical analysis based on the network calculus theory can be used to provide the deterministic performance bounds. Through the extensive simulations, the proposed architecture and related algorithm can be proved efficient and reliable in obtaining the fairness and adaption, which can be the foundation of resource management in order to support all kinds of network services.

References

[1] Yunkai Zhou, Harish Sethu, *"On Achieving Fairness in the Joint Allocation of Processing and Bandwidth Resources"*, IWQoS 2003, LNCS 2707, 2003, Page(s): 97 - 114

[2] Yunkai Zhou, Sethu, H., *"On achieving fairness in the joint allocation of processing and bandwidth resources: principles and algorithms"*, IEEE/ACM Transactions on Networking, Volume 13, Issue 5, Oct. 2005, Page(s):1054 - 1067

[3] A. Parekh, R. Gallager, *"A generalized processor sharing approach to flow control in integrated services networks: The single node case*, IEEE/ACM Trans. Networking, vol.1, June 1993, Page(s): 344-357

[4] J. Golestani, *"A self-clocked fair queuing scheme for broadband applications*, in Proc. of IEEE INFOCOM, 1994, Page(s): 636 -646

[5] J. C. R. Bennett and H. Zhang, *"WF2Q: Worst-case fair weighted fair queuing*, in Proc. of IEEE INFOCOM, 1996, Page(s): 120 -128

[6] D. Stiliadis, A. Verma, *"Rate-proportional servers: A design methodology for fair queuing algorithms*, IEEE/ACM Trans. Networking, vol.6, 1998, Page(s): 164-174

[7] D. Stiliadis, A. Verma, *"Efficient fair queuing algorithms for packet-switched networks*, IEEE/ACM Trans. Networking, vol. 6, 1998, Page(s): 175-185

[8] J. C. R. Bennet, H. Zhang, *"Hierarchical packet fair queuing algorithms*, IEEE/ACM Trans. Networking, vol. 5, 1997, Page(s): 675-689

[9] M. Shreedhar, G. Varghese, *"Efficient fair queuing using deficit round-robin*, IEEE/ACM Trans. Networking, vol. 4, 1996, Page(s): 375-385

[10] Luciano Lenzini, Enzo Mingozzi, and Giovanni Stea, *"Tradeoffs Between Low Complexity, Low Latency,and Fairness With Deficit Round-Robin Schedulers"*, IEEE/ACM Transaction on networking, vol. 12, no.4, August 2004

[11] S.-C. Tsao, Y.-D. Lin, *"Pre-order deficit round-robin: A new scheduling algorithm for packet switched networks*, Comput. Netw., vol. 35, 2001, Page(s): 287-305

[12] Yu Hua, Chanle Wu, and Mengxiao Wu, *"Scheduling Scheme with Fairness and Adaptation in the Joint Allocation of Herogeneous Resources"*, in Proc. of the Advanced Parallel Processing Technologies (APPT), 2005, Page(s): 151-163

[13] Liang Xu; Xuemin Shen; Mark, J.W., *"Fair resource allocation with guaranteed statistical QoS for multimedia traffic in wideband CDMA cellular network"*, IEEE Transactions on Mobile Computing, Volume 4, Issue 2, March-April 2005, Page(s):166-177

[14] Drougas, Y.; Kalogeraki, V., *"A fair resource allocation algorithm for peer-to-peer overlays"*, in Proc. of INFOCOM, IEEE Volume 4, 13-17 March 2005, Page(s): 2853-2858

[15] R. L. Cruz, *"A calculus for network delay, Part I: Network elements in isolation,* IEEE Trans. Inform. Theory, vol.37, Jan.1991, Page(s): 114-131

[16] Krishna Pandit, Jens Schmitt and Ralf Steinmetz, *"Network Calculus meets Queueing Theory -A Simulation Based Approach to Bounded Queues",* in Proc. of IWQOS, June 2004, Page(s): 114-120

[17] NS-2: http://www.isi.edu/nsnam/ns/

[18] G. Chuanxiong, *"SRR: An O(1) time complexity packet scheduler for flows in multi-service packet networks*, in Proc. of ACM SIGCOMM, 2001, Page(s): 211-222

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.