

论文题目：基于输入排队的可扩展交换结构调度算法的研究

摘 要

交换结构(switch fabric)作为交换机和路由器的核心,如何提高其交换容量可扩展性和服务质量可预测性,是近十年来网络研究的一个热点和难点问题。一个典型的交换结构由三部分组成:输入端口、输出端口和交换内核。为了避免来自不同输入端口的信元同时发往同一个输出端口,需要在输出端口或者输入端口设置缓冲区,从而形成输出排队和输入排队两大基本交换结构。尽管输出排队型交换结构可以提供良好的服务质量保证(100%的吞吐率,有界的延迟,带宽公平性等),然而其存储器的带宽却需要所有输入端口带宽总和,这极大限制了其扩展性。与之相比,输入排队型交换结构允许内部存储器的带宽工作在线速,其良好的可扩展性使其成为高性能路由器的主流交换结构。由于交换结构的调度算法负责将输入端口的信元通过交换内核发送至输出端口,所以它在提高交换设备的利用率及其服务质量保证方面起着关键性作用。本文从交换容量的可扩展性及服务质量可预测性的角度出发,研究了基于输入排队的不同类型交换结构下的调度算法设计。

目前,核心交换机/路由器的主流交换结构一般采用交叉开关(crossbar)以保证交换内核无阻塞,并采用集中式调度器调度定长信元通过交叉开关。对于该交换结构,具有较强理论意义的一类算法为最大权重匹配算法,已证明对于任意容许的流量,均能达到100%的吞吐率,并且平均延迟有界,然而其算法的复杂度高达 $O(N^3)$,本文从局部搜索的角度研究了最大权重匹配的近似算法,结合局部搜索的可并行计算的特点,提出了一种并行随机调度算法及一种并行确定性调度算法,并且证明了算法的稳定性,与已有近似算法相比,具有更低的平均延迟。

缓冲交叉开关型交换结构由于具有分布式存储及分布式调度的特点,是构建特比特级(Terabit)路由器的一种理想选择。由于轮转型调度算法易于硬件实现,具有较高的应用价值,从而得到了广泛的研究。现有轮转型算法在调度均匀流量时具有逼近100%的吞吐率,然而对于非均匀的流量,现有轮转型算法的吞吐率却明显下降。为解决此问题,与当前的单轮转指针不同,本文提出了一类双轮转指针的调度算法,即在每个输入调度器均设置了主指针与辅助指针,主指针对应的队列具有最高的调度优先级,算法可以根据各个队列的状态来动态决定何时更新主指针,当主指针对应的队列被流控机制阻塞时,将根据辅助指针依次公平服务其他队列。仿真实验表明,对于每个交叉点缓冲区仅有一个信元的交换结构,

基于双指针的调度算法可以显著提高该交换结构在已知多种非均匀流量下的性能。

对于缓冲交叉开关型交换结构，一般采用基于份额的流控机制，在这种方式下，为了确保输入和输出端口都可以工作保持(work-conserving)，每一个交叉点缓冲区大小至少需要线速乘以交换结构内部环路延迟，对于特比特级、多机柜的交换机，其交叉点缓冲区的需求必然很大，从而给实现带来困难。本文从均匀交换的角度研究了该交换结构下的服务质量保证问题，提出了一种新型的支持流一级均匀交换的内核，该交换结构采用基于位率信息的流控机制，允许任意大小的交换内核至线卡的往返延迟，并且在容许的流量下，每个交叉点缓冲区的容量仅需四个信元即可保证 100% 的吞吐率，并且没有信元丢失。

关键词： 路由器；交换结构；调度算法；虚拟输出排队；服务质量；均匀性

Research on the Design of Scheduling Algorithms for Input Queued Scalable Switches

Zheng Yanfeng

Directed By Gao Wen

Switch fabric is a core element of high performance switch and router. How to increase its switching capacity and the predictability of quality of service is a hot and hard topic of network research over the past decade. A typical switch fabric consists of three elements: input ports, output ports and switch core. To avoid cells from different input ports being sent to the same output port at the same time, it is necessary to place buffers at output ports or input ports, and two buffering schemes are prevalent: output queueing (OQ) and input queueing (IQ). Although the OQ switches provide good quality of service guarantee (100% throughput, bounded delay, bandwidth fairness etc.), the memory bandwidth requires the sum of all the input ports bandwidth. Therefore the scalability of OQ switches is greatly limited. Compared with the OQ switches, the IQ switches have a memory bandwidth requirement comparable to the line rate, and such good scalability makes IQ switch fabrics becoming the prevalent choice of high performance router. Since the scheduling algorithm is responsible to transfer the cells from the input ports to the output ports across the switch core, it plays the key role in improving the switch devices utilization and the quality of service guarantee. From the view points of the scalability of switching capacity and the predictability of quality of service, this thesis studied the design of scheduling algorithms for some types of scalable switch fabrics.

At present, crossbar is often employed to construct the switch core because of its nonblocking property, and one centralized scheduler is used to schedule the cells across the crossbar. For this switch fabric, maximum weighted matching

(abbr. MWM) algorithms have been proved to achieve 100% throughput and the bounded average delay. However such type of algorithms has high time complexity $O(N^3)$. This thesis studied MWM approximations based on local search technique. Combined with the parallel computing feature of local search, we presented a randomized scheduling algorithm and a deterministic scheduling algorithm. Furthermore, we proved the stability of the presented algorithms. Compared with the existing approximations, our algorithms have lower average delay.

The buffered crossbar switch fabric is an ideal choice to build the terabit-capacity router because of its features of distributed storage and distributed scheduling. Since the round-robin algorithms are easy to be implemented by hardware, such a type of algorithm has been widely studied. Although the previously proposed round-robin algorithms provide nearly 100% throughput under uniform traffic, the throughput of such algorithms will drop down under nonuniform traffic patterns. In order to solve this problem, we presented a class of dual round-robin algorithms which is different from the single round-robin pointer usage. For our algorithms, each input scheduler is associated with a primary pointer and a secondary pointer. The input queue pointed to by the primary pointer has the highest priority to be scheduled; the input scheduler decides when to update the primary pointer according to the status of input queue. When the input queue pointed to by the primary pointer is blocked by the flow control, the secondary pointer is used to service other queues in a fair way. The extensive simulations show that the performance of buffered crossbar switch under nonuniform traffic is greatly improved by introducing the dual round-robin pointers method.

For buffered crossbar switch fabric, the credit based flow control is often used. Under such mechanism, in order to keep work-conserving of input and output port, the capacity of each crosspoint buffer should be at least equal to the product of the line rate and the fabric-internal round-trip latency. Such large crosspoint buffer requirement will be a hurdle to implementing multiterabit switches. This thesis studied the quality of service of buffered crossbar switches from the point of view of smooth switching. A new type of switch core which supports the flow level smooth switching is presented. Such switch fabric adopts rate based flow control method, which permits arbitrary fabric-internal latency. Under admissible traffic, it provides 100% throughput and no cell loss with only four cells at each crosspoint buffer.

Keywords: Router, Switch Fabric, Scheduling Algorithm, Virtual Output Queuing, Quality of Service, Smoothness