# 快速自适应调频机制及其在 NetFPGA 上的实现

汪 漪[1) ] 孟 玮[1)] 胡成臣[2)] 贺可强[1)] 刘 斌[1)]

[1)](清华大学计算机科学与技术系 北京 100084)

[2)](西安交通大学计算机科学与技术系 西安 710049)

**摘 要** 降低网络设备能耗已经成为当前研究的热点.文中,作者设计了一种能够实时适应流量变化的动态自适应频率调整方法 FASS,它能根据处理模块的负载,实时地调整模块的工作频率,从而有效降低模块的能耗.同时,作者通过修改 NetFPGA 的参考路由器设计,将 FASS 添加到数据包处理模块中来验证 FASS 的性能.马尔可夫模型分析结果和实际实验的测试结果表明,在仅增加可容忍的延迟的情况下,FASS 在多种不同的负载情况下,都能有效地降低模块功耗.另外,作者的工作说明 FASS 可以应用于实际的物理设备中以实现节能.

**关键词** 自适应频率调整;双门限多阈值;NetFPGA;节能;路由器;绿色计算;绿色网络

**中图法分类号** TP393 **DOI 号**: 10.3724/SP.J.1016.2012.01286

# Fast Adaptive Speed Scaling Mechanism and Implementation on NetFPGA Platform

WANG Yi[1)] MENG Wei[1)] HU Cheng-Chen[2)] HE Ke-Qiang[1)] LIU Bin[1)]

[1)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

[2)](*Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049)

**Abstract** The concept of energy-efficient networking has been a hot research topic in the past few years, gaining increasing popularity. In this paper, we design a fast adaptive speed scaling mechanism to reduce power consumption of network devices. The operation frequency of components in a network device is adjusted dynamically to different levels according to the real time workload. We implement a prototype of this mechanism in the data path of a general IPv4 router based on a real hardware platform—— NetFPGA. The theoretical analysis and experimental results show excellent energy savings at the cost of a tolerable latency, under various ranges of traffic loads. Our work indicates the feasibility and possibility of deploying the mechanism into real network devices for energy saving.

**Keywords** adaptive speed scaling; multi-dual-threshold; NetFPGA; power saving; router; green computing; green networking

## 1 引 言

根据 Uclue 组织的研究结果①,2007 年 Internet 消耗的电量占美国总电量消耗的 9.4%,占全球总电量的 5.3%.当前 Internet 骨干网的带宽平均利用率为 30%,带宽最高平均利用率小于 45%②,而网络设备却都满负荷运行以应对网络流量突发,这

收稿日期:2011-08-07;最终修改稿收到日期:2012-02-15.本课题得到国家自然科学基金(61073171,60873250)、高等学校博士学科点专项科研基金(20100002110051)、清华大学自主科研计划资助.**汪 漪**,男,1983 年生,博士研究生,主要研究方向为高速路由器设计与实现、数据包转发和内容标记网络. E-mail: wy@ieee.org. **孟 玮**,男,1990 年生,本科生,主要研究方向为网络系统与安全. **胡成臣**,男,1981 年生,博士,副教授,主要研究方向为网络系统、网络管理和测量. **贺可强**,男,1987 年生,硕士研究生,主要研究方向为以太网交换机设计与实现. **刘 斌**,男,1964 年生,博士,教授,主要研究领域为网络系统、路由器体系结构、网络处理器和内容标记网络.

① Energy Use of Internet. http://uclue.com/index.php?xq=724, 2007

② http://arstechnica.com/old/content/2008/09/what-exaflood-net-backboneshows-no-signs-of-osteoporosis.ars, 2008

就造成大量的电能被空闲网络基础设施所消耗与浪费[1]. 为降低网络设备的能耗，学术界和工业界提出许多能量管理的方法，大致可分为两类：(1) 协议方式，通过创建、修改协议来实现全网协同(例如流量聚合)来降低功耗；(2) 设备方式，通过重新设计设备体系结构和优化处理单元的运行模式来实现节能. 设备休眠法和频率调整法是降低处理模块功耗的两种基本方法. 设备休眠法通过将网络设备在运行状态和休眠状态之间不断地切换来实现节能，其中什么时候切换，切换多长时间是该方法的核心问题[2]. 设备休眠法在流量变化不大，或可提前预知的情况下，有较好的效果，但对于以突发为基本特点的网络流，则会造成网络性能的急剧下降[3]. 频率调整法根据实时的负载情况，通过动态地调整物理器件的时钟频率来实现节能，但支持调频的物理器件较少，并且传统的频率调整方法只能适应较慢的流量变化，整体节能效果较差. 通常情况下，无论外部到达的流量如何变化，实际的路由器、交换机等网络设备都工作在最高频率，进行数据包的存储与转发等操作. 在过去几年中，动态链路关闭机制[2]和自适应链路速率调整机制[4]被提出并应用于以太网设备的物理链路芯片和接口中，但设备的路由查找、转发、背板交换网络等其余模块仍旧满负荷工作，使得基于这种机制不能达到理想的节能效果.

本文中，我们设计了一种能够实时适应流量变化的动态自适应频率调整方法(Fast Adaptive Speed Scaling，FASS)，并将 FASS 应用于 NetFPGA 平台①的参考路由器数据平面的各个数据包处理模块中来进行验证. NetFPGA 板卡具有 4 个千兆以太网接口，核心处理部件采用可编程的 Xilinx Virtex II FPGA，通过 PCI 接口与台式电脑主板相连. NetFPGA 是一个基于 Linux 的开放性平台，它为研究人员提供了原型快速开发与实现的硬件平台，开源代码和脚本帮助研究人员更好地使用、修改参考设计，甚至创建全新的系统和应用.

FASS 的设计目标是能够智能地根据工作模块的负载，快速、动态地调整工作频率，在不显著增加等待、处理时延的约束下，最大程度地降低模块的能耗. 要实现以上目标，FASS 需要攻克以下两个难点：

(1) 频率切换策略. 频率切换策略应该保证在时延、资源开销的约束下，及时响应输入流量的变化，最优化模块的操作频率，从而降低模块的能耗；

(2) 快速切换机制与物理实现. 不仅能适用于实际物理设备，而且具有占有较小系统资源和支持快速切换速度机制的特性，为 FASS 适用于流量快速变化的网络设备提供基础支持.

本文中，我们主要有以下 4 点贡献：

(1) 提出一种渐进的双门限多阈值的频率实时自适应方法(FASS)，它能根据等待处理的队列长度，平稳地自动在多个模块工作频率之间切换，有效降低网络设备中物理模块的能耗.

(2) 采用异步时钟实现数据包处理模块之间的协同工作，并设计和实现了仅需要一个时钟周期延迟的快速频率切换机制.

(3) 通过修改 NetFPGA 平台上开源 IPv4 参考路由器，添加 FASS 到数据包处理的模块中，验证 FASS 的节能效果.

(4) 使用“SmartBits 600 网络性能分析仪”②来实际测试 FASS 引起的数据包处理时延的增加量，实际实验结果表明 FASS 能够在只增加较小时延的情况下，有效降低能耗.

本文第 2 节对动态自适应频率调整机制进行阐述，并在第 3 节中应用马尔可夫链对 FASS 进行理论分析；第 4 节描述如何在 NetFPGA 参考路由器中实现动态自适应频率调整机制，其实际测试结果在第 5 节中给出；动态频率调节的相关研究工作在第 6 节中论述，并在第 7 节中对本文工作进行总结.

## 2　动态自适应频率调整机制

### 2.1　FASS 的性能指标

网络设备的动态自适应频率调整机制在有效降低工作模块能耗的同时，会造成单个工作模块和系统整体时延的增加. 所以我们基于如下两个关键性能指标对 FASS 进行分析和验证：

(1) 节能效果. 由于设备的能耗与速率基本成线性关系[5]，我们使用相对速率(实际速率与最高速率的比值)来间接衡量节能的效果. 特别地，我们采用速率降低的平均值而不是瞬时值来更宏观、准确地描述节能效果.

(2) 数据包时延. 一般而言，系统、模块工作在低频模式相比工作在高频模式会带来更大的时延. 更进一步，由于频率切换造成的数据包时延的抖动可能会对基于数据包时延进行路由选择的路由协议带来负面影响. 因此，自适应频率调节方法必须使得其引起的时延增加限制在一定范围. 总体而言，数据包时延增量主要由队列排队时延引起. 在本文中，我

① NetFPGA Program. http://www.netfpga.org/
② Spirent smartbits. http://www.spirent.com/Solutions-Directory/Smartbits.aspx

们采用队列排队时延作为时延的度量.

一个自适应频率调节机制需要在节能效果和数据包时延之间寻求平衡,即在特定的时延约束和输入流量动态变化的情况下,最大程度地节约模块能耗.在接下来的章节中,我们用上面定义的两个指标对提出的 FASS 进行理论分析和实验测试.结果表明 FASS 在增加可容忍的延迟的情况下,能实现较佳的节能效果.

### 2.2 双门限多阈值机制

现今的网络设备大多使用缓存应对短时间的突发流量,并在各个功能部件之间对任务队列进行缓冲.在图 1 中,以太网帧处理模块和数据包处理模块之间就通过缓存队列进行连接.直观上,缓存的占用率(队列长度)可以反映设备的实时负载.在我们的 FASS 机制中,缓存的实际占用情况是进行频率切换的主要依据.我们假设硬件支持工作在多个频率下,并设置了一系列阈值来近似衡量设备的工作负载.

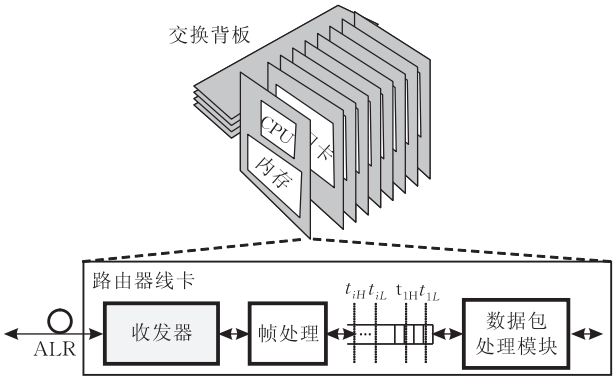
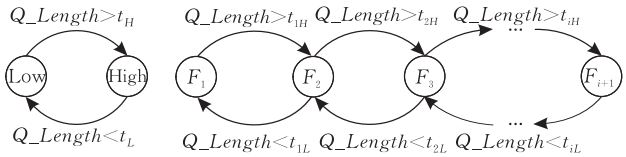



图 1 路由器线卡中各个模块之间通存队列相连

实时地根据输入队列长度来进行频率的动态调整是一种有效的节约能耗的方法,它使得模块在线路负载较轻的时候,工作在较低频率,实现按需工作,从而达到较低功耗的目的.当仅使用一个队列长度阈值为调频依据时,如果队列长度恰好在阈值附近抖动,就会带来频率切换的抖动,造成额外的切换时延和功耗.双阈值机制,即一个高阈值 $t_H$ 和一个低阈值 $t_L$,正是为避免单阈值带来的切换抖动的发生而提出.频率切换的状态图如图 2(a)所示,当模块工作在低频模式,队列长度由小到大增长超过 $t_H$ 时,工作频率被切换到高频模式;当模块工作在高频模式时,队列长度由大到小减小到低于 $t_L$ 时,频率被切换到低频模式.

双阈值机制的节能效果与阈值的取值密切相关,当阈值与负载情况较符合时,可以有效降低能耗,而当两者不相符合时,特别是在 $t_L$ < 负载 < $t_H$ 时,

(1)会带来时延的急剧增加(模块工作在低频




(a) 双阈值机制 (b) 多双阈值机制

图 2 频率切换状态示意图

模式,且队列长度始终不超过 $t_H$).在这种情况下,模块一直工作在低频模式,而队列的长度又相对较长,造成时延的急剧增加.

(2)不能有效地降低功耗.当模块工作在高频模式,且队列长度始终不低于阈值 $t_L$,在这种情况下,模块一直工作在高频模式,没有达到降低功耗的目的.

(3)为最大程度地降低功耗,双阈值机制一般设置低频模式时的处理频率较低,往往比高频模式时的处理频率小一个数量级,例如 ALR[4] 中的 100 Mbps 和 1 Gbps.这就带来另外一个问题:数据包时延的抖动过于大,使得一些以数据包传送时延为依据的拥塞控制机制失效(负载轻时,时延大;负载重时,时延反而轻).

为解决双阈值机制中的几个基本问题,我们提出了双门限多阈值机制,其频率切换的状态转移图如图 2(b)所示.其中,规定 $t_{iH}<t_{(i+1)L}<t_{(i+1)H}$.两个相邻的频率之间的切换,仍然采用双阈值机制来确定,模块工作频率只能在相邻的频率之间切换,不能越级切换.相邻的频率之间,有 $f_i=\frac{1}{2}f_i+1$.相比双阈值机制,$t_{iH}-t_{iL}$ 小于 $t_H-t_L$,从而使得负载落于区间 $t_{iH}-t_{iL}$ 的概率降低,且只要一个较小幅度的流量突发,就能打破原来的僵局,实现频率的切换,从而实现较小的时延和减少更多的能耗的目标.另一方面,由于存在多个频率,减小了相邻频率之间的差值,使得模块工作频率切换带来的数据包的时延抖动较小,更符合网络实际的负载情况.

## 3 理论分析

在这一节中,我们给出双门限多阈值策略的数学定义,并用马尔卡夫链进行性能分析.之后,我们采用上一节中定义的能耗与时延的标准对速率和阈值的设置进行研究.

### 3.1 双门限多阈值策略

首先,假设数据包的到达和服务时间服从泊松分布.双门限多阈值策略可以使用单一队列、单一服

务排队论模型进行建模，其服务速率取决于服务所处的工作状态，双门限被用于避免潜在的频率抖动及其造成的系统不稳定.

假设 $\lambda$ 为数据包的到达速率，且系统能够工作在 $M+1$ 种不同的速率，从 $\mu_1$ 到 $\mu_{M+1}$（$\mu_j<\mu_{j+1}$，$1\leq j\leq M+1$）. 相应地，利用率可以表示为 $\rho_j=\lambda/\mu_j$. 在双门限多阈值策略中，我们定义了 $2M$ 个阈值用以进行速率切换. 其中，$M$ 个低阈值表示为 $T_{jL}$（$1\leq j\leq M$），$M$ 个高阈值表示为 $T_{jH}$（$1\leq j\leq M$），并且有 $T_{jL}<T_{jH}<T_{(j+1)L}$，$1\leq j\leq M-1$. 当缓存占用降低至低阈值 $T_{jL}$ 以下时，会引起服务速率由 $\mu_{j+1}$ 降至 $\mu_j$；而当缓存占用升至高阈值 $T_{jH}$ 以上时，服务速率将由 $\mu_j$ 提高至 $\mu_{j+1}$.

接下来，我们将双门限多阈值策略建模成马尔卡夫链以对其进行分析，并给出系统（缓存）中有 $n$ 个客户（数据包）的稳态概率 $P_n$. 需要特别注意的是，我们使用符号 $\pi_n$ 来表示马尔卡夫模型中状态 $n$ 的稳态概率，这不同于 $P_n$.

### 3.2　稳态概率

图 3 中显示了双门限多阈值调频策略的马尔卡夫链，其中位于顶部的一行表示与其下方状态相对应的缓存占用率（队列长度）. 所使用的参数和变量定义如下：$\lambda$ 为数据包到达速率；$M$ 为高/低阈值数量；$\mu_i$ 为第 $i$ 档服务速率（$1\leq i\leq M+1$）；$\rho_i=\lambda/\mu_i$ 为 $\mu_i$ 对应的利用率；$k_{2i-1}$ 表示缓存占用的低阈值 $T_{iL}$（$1\leq i\leq M$）；$k_{2i}$ 表示缓存占用的高阈值 $T_{iH}$（$1\leq i\leq M$）；$\pi_n$ 为状态 $n$ 的稳态概率. 对处于 $k_{2i+1}$ 和 $k_{2i+2}$（$1\leq i<M$）之间的缓存占用，例如 $k_{2i+1}+j$，对应于两个状态，则用 $\pi_{k_{2i+1}+2j}$ 表示位于上链状态的稳态概率，用 $\pi_{k_{2i+1}+2j+1}$ 表示位于下链状态的稳态概率；$P_n$ 对应于队列长度为 $n$ 的稳态概率. 对于只对应于一个状态的情况下，$P_n=\pi_n$；对于对应两个状态的情况下，$P_n$ 则为两个对应状态的稳态概率的和，例如 $P_{k_1}=\pi_{k_1}+\pi_{k_1+1}$；$T_i$ 是对应于服务速率 $\mu_i$ 的时间占用率.

特别地，$k_0$ 表示缓存占用为 0 的情况，而用 $P_{k_{2i}}$ 代表缓存占用为高阈值 $TH_i$（$1\leq i\leq M$）的情况下的稳态概率.

在给出上述定义之后，其最终的稳态概率、时间占用率、相对速率的数学表达式如下所示，具体推导过程详见文献[6].

队列长度为 $n$ 的稳态概率为

$$P_n=\begin{cases}\rho_{i+1}^{n-k_{2i}}P_{k_{2i}}, & k_{2i}\leq n<k_{2i+1},0\leq i<M\\ \pi_{k_{2i+1}+2(n-k_{2i+1})}+\pi_{k_{2i+1}+2(n-k_{2i+1})+1}, & k_{2i+1}\leq n<k_{2i+2},0\leq i<M\\ \rho_{M+1}^{n-k_{2M}}P_{k_{2M}}, & n\geq k_{2M}\end{cases}\tag{1}$$

其中

$$P_{k_{2i+2}}=\frac{(1-\rho_{i+2}^{k_{2i+2}-k_{2i+1}+1})(1-\rho_{i+1})\rho_{i+2}\rho_{i+1}^{k_{2i+2}-1}}{(1-\rho_{i+2})(1-\rho_{i+1}^{k_{2i+2}-k_{2i+1}+1})}P_{k_{2i}},\quad 0\leq i<M.$$

由于服务速率大小不确定，因而无法直接推出 $P_0$ 的表达式. 然而，由于 $P_n$ 从 0 到无穷的积分等于 1，只要给定了 $M$，可以近似求得 $P_0$ 的值.

不同服务速率的时间占用率可以作为衡量节能效果的标度，对于 $\mu_i$ 的时间占用率 $T_i$ 可以由如下表达式描述：

$$T_i=\begin{cases}\sum_{n=0}^{k_1-1}\pi_n+\sum_{j=0}^{k_2-k_1-1}\pi_{k_1+2j}, & i=1\\ \sum_{j=0}^{k_{2i-2}-k_{2i-3}-1}\pi_{k_{2i-3}+2j+1}+\sum_{n=k_{2i-2}}^{k_{2i-1}-1}\pi_n+\sum_{j=0}^{k_{2i}-k_{2i-1}-1}\pi_{k_{2i-1}+2j}, & 2\leq i\leq M\\ \sum_{j=0}^{k_{2M}-k_{2M-1}-1}\pi_{k_{2M-1}+2j+1}+\sum_{n=k_{2M}}^{\infty}\pi_n, & i=M+1\end{cases}\tag{2}$$

因而，相对速率可以表达为

$$\text{相对速率}=\sum_{i=1}^{M+1}T_i\frac{\mu_i}{\mu_{M+1}}\Big/\sum_{i=1}^{M+1}T_i\tag{3}$$

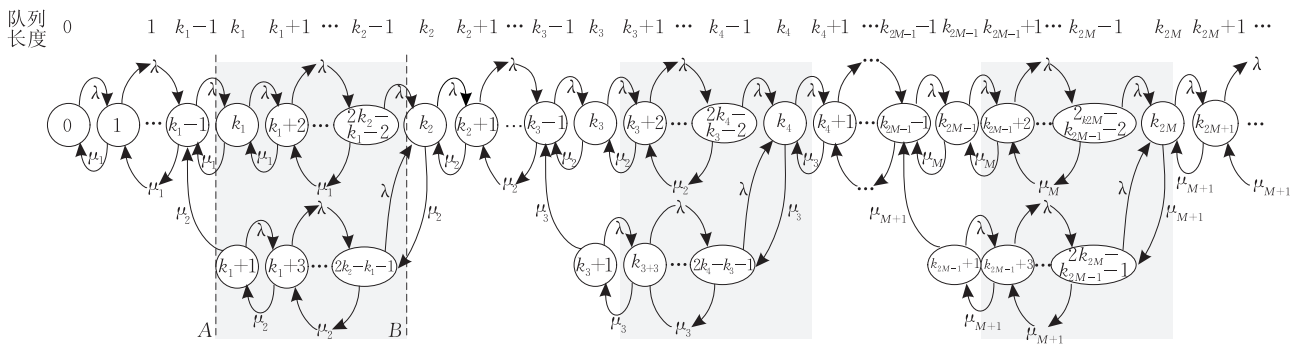



图 3　多双阈值的马尔可夫链

队列排队时延可以表达为

$$Delay=\sum_{i=0}^{M-1}\sum_{n=k_{2i}}^{k_{2i+1}-1}\pi_n\frac{n}{\mu_{i+1}}+\sum_{n=k_{2M}}^{\infty}\pi_n\frac{n}{\mu_{M+1}}+\sum_{i=0}^{M-1}\sum_{n=k_{2i+1}}^{k_{2i+2}-1}n\left(\frac{\pi_{2n-k_{2i+1}}}{\mu_{i+1}}+\frac{\pi_{2n-k_{2i+1}+1}}{\mu_{i+2}}\right)\quad(4)$$

### 3.3 速率和阈值设置

在得到上述表达式(1)～(4)后,我们在本节中分析速率和阈值分布对系统性能的影响.我们选取如下3种有代表性的速率和阈值分布进行分析:

(1)均匀速率、均匀阈值,$N$个不同的速率在$1/N$最大速率($\frac{1}{N}$Gbps)和最大速率(1Gbps)间均匀分布,$N-1$个高阈值亦按均匀分布,表示为$URUT-N$.

(2)指数速率、均匀阈值,($L=\log_2 N+1$)个速率在$1/N$最大速率($\frac{1}{N}$Gbps)和最大速率(1Gbps)间指数分布,$\log_2 N$个高阈值按均匀分布,表示为$ERUT-L$.

(3)指数速率、指数阈值,($L=\log_2 N+1$)个速率在$1/N$最大速率($\frac{1}{N}$Gbps)和最大速率(1Gbps)间按指数分布,$\log_2 N$个高阈值亦按指数分布,表示为$ERET-L$.

在这3种设置中,低速率的分布与高速率相同,并假设每对高低速率的差相等.对于指数分布,使用2作为幂.在接下来的理论估计中,缓存大小设置为32KB.

除了分析速率及阈值的分布外,高低阈值之间的差异(亦即频率切换的缓冲区大小)也需要考虑.为此,我们在上面3种分布的基础上额外独立设置了两个关于高低阈值差异的配置.配置$A$的高低阈值差接近于0,亦即$(TL_i\approx TH_i)\wedge(TL_i<TH_i,1\leq i\leq M)$,这种情况是对单门限策略的趋近;配置$B$的高低阈值差接近于最大可能值,亦即$(TL_i\simeq TH_{i-1})\wedge(TH_{i-1}<TL_i)$,$1<i\leq M$,这模拟了较大频率切换缓冲区的情况.

图4显示了配置$A$和配置$B$的3种分布在不同系统负载情况下的平均队列时延,图5显示了配置$A$和配置$B$的3种分布在不同系统负载情况下的平均相对速率,其中$N$的值均为16.如图4所示,在3种分布中,$ERET$-5的队列时延最低,而$ERUT$-5的队列时延则比另外两个分布高出很多,这种情况在系统利用率不高的时候尤为明显.这说明阈值的分布应该与所选取的速率分布相一致或匹配.对于平均相对速率,3种分布的效果非常接近,且均与系统的利用率成比例.从图5中可以看出,与配置$A$相比,配置$B$的相对速率更加理想,更接近于最优值.对于队列时延,配置$B$也略低于配置$A$,这是由于频率切换的缓冲区加大了(马尔卡夫模型中的单链变短),从平均意义而言,系统有更多的机会工作在相对较高的频率下.另一方面,由于缓冲区变大,系统进行频率切换的次数会减少,这保证了系统的稳定,也减少了频率切换所带来的开销.总体而言,上述结果说明双门限策略比单门限策略的性能更优,且性能会随着高低阈值的差增大而提升.

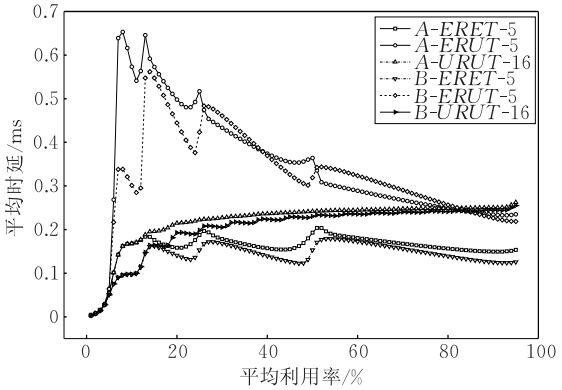



图4 不同分布、阈值在不同负载下的平均时延(最大负载为1Gbps)

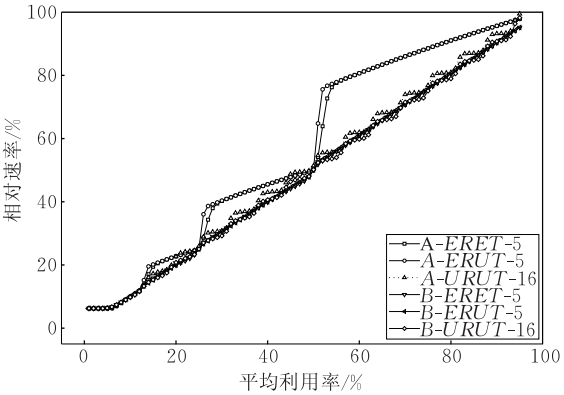



图5 不同分布、阈值在不同负载下的相对速率(最大负载为1Gbps)

从图5中,我们可以注意到,在低频时系统所能降低的频率近似于$1-1/N$,这说明频率划分越细(即$N$越大),或是最低频率越低,在低频时所能节约的能量越多.而对于系统利用率高于$1/N$时,$N$对相对速率的大小没有影响.因而在实际应用中,过度追求频率的细分和过低的频率是没有必要的.

综上所述,频率和阈值的分布对系统性能有着非常大的影响,对于所选取的3种频率和阈值的分布,第3种指数频率指数阈值的性能最佳.特别地,高低阈值的差异应该设置地尽可能大,以降低队列

延迟和增强系统稳定性.另外,阈值的分布应该与频率的分布相匹配,而频率划分的粒度和最低频率的大小仅对系统利用率低时的功耗有着有限的影响.而由于阈值的选取是基于缓存的大小的,一般而言缓存取值越小,系统可以在队列更短的时候切换到高频状态,有利于减少因调频带来的额外队列时延.因此,在理论上使用 FASS 进行路由器数据通路的节能是可行的.为了进一步验证我们的分析,我们在 NetFPGA 平台的参考路由器上实现了 FASS,并进行了实际流量的测试,这将在接下来的章节中讨论.

## 4　FASS 在 NetFPGA 参考路由器中的实现

### 4.1　NetFPGA 参考路由器

NetFPGA 是美国斯坦福大学开发的,具有千兆速率(以太网 RJ45 接口)的可编程硬件平台.NetFPGA 是基于 Linux 的开源项目,所有参考的路由器、交换机、网卡等设计与实现都面向研究人员和学生开发,为大家提供了一个实用、高效的网络系统开发平台.研究人员可以通过修改参考设计中的部分某块或者添加自定义模块来快速实现、验证自己的想法.

当前有两个版本的 NetFPGA 板卡:NetFPGA-1G 和 NetFPGA-10G,FASS 通过修改 NetFPGA-1G 的参考路由器进行实际的性能测试.NetFPGA 核心处理器是一块 Xilinx Virtex II Pro 50 FPGA[7],工作频率可配置为 125 MHz 或 62.5 MHz.另外,两块 SRAM 协同核心 FPGA 实现数据包和其它数据的存储.在我们的实现中,FPGA 和 SRAM 的工作频率设置为 125M.

NetFPGA-1G 的参考路由器体系架构如图 6 所示,共有 8 对接收和发送队列,其中 4 对队列对应 4 个千兆以太网 RJ45 接口,其余 4 对队列对应与其相连的主机的 CPU-DMA 接口.图 6 中的时钟适配器是 FASS 添加的模块,用于动态调整各个模块的工作频率.在用户数据通路上的流水架构采用 64 比特的总线宽度,在 125 MHz 时钟下,最高支持 8 Gbps 的带宽[8].所有的内部模块采用标准的先进先出的请求、授权协议.使用者可以通过额外的队列和模块到用户数据通路上,实现需求的功能.

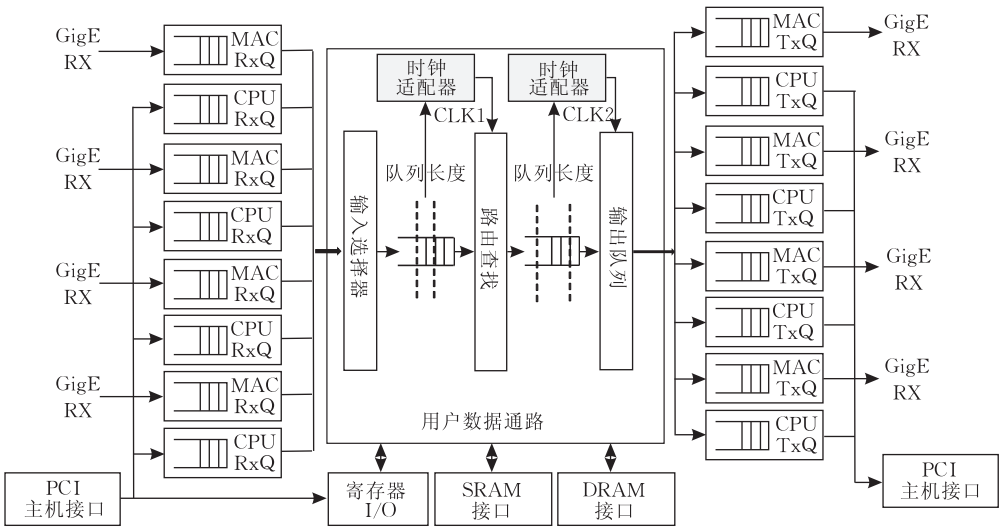



图 6　NetFPGA 参考路由器数据通路流水架构

### 4.2　具有 FASS 功能的原理路由器实现

具有 FASS 功能的原理路由器是以 NetFPGA 的参考路由器为基础架构,添加少量频率调节模块和队列到数据通路上实现的,如图 6 中阴影部分所示.原有的模块之间的缓冲队列都是同步队列,为实现各个模块异步的动态工作频率调整,我们使用异步队列替换所有原先的同步队列以支持前后级模块工作在不同的频率.所有异步队列的数据位宽都与之前的同步队列的数据位宽相同,为 64 比特,保持了原先数据通路最高带宽为 8 Gbps 的特性.频率调整模块根据实际等待队列的长度,动态地调整模块的工作频率、与之对应的输入队列的输出时钟频率(读时钟频率)和输出队列的输入时钟频率(写时钟频率),如图 7 所示.相邻的模块可以根据自身处理速率和等待队列长度的不同,动态地调整各自的工作频率,从而实现各个模块的异步调频,实现降低模块功耗目标.

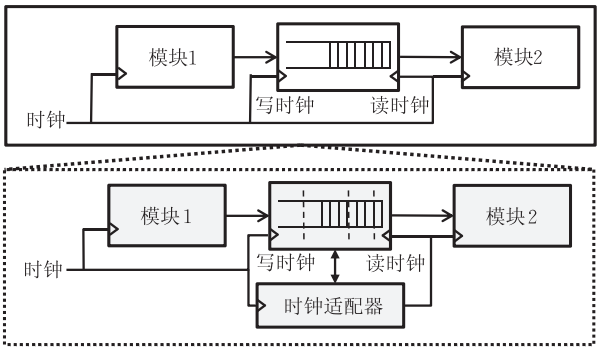



图 7 动态频率调整模型

在我们目前实现的原型系统中,频率适配器被设置在“输入选择器”模块、“路由查找”模块以及“输出队列”中的子模块中.频率适配器具有 6 档频率,分别是 125 MHz、62.5 MHz、31.25 MHz,15.625 MHz,7.813 MHz 和 3.096 MHz,相邻两档频率时间是二倍频关系.由于 SRAM 的工作频率是与 FPGA 芯片的工作频率一致的,我们在原型系统中没有实现对 SRAM 频率的动态调整.

NetFPGA 的寄存器接口允许运行在主机上的软件通过 PCI 接口与 NetFPGA 接口卡通信,发送数据到寄存器(指令等)或查询寄存器的数据.为测量采用 FASS 路由器原型系统的性能,我们添加频率计数寄存器到 NetFPGA 的寄存器组中,实时记录各个模块在不同频率下运行的次数(时间).

## 5 实验结果

由于路由器的基本功能是快速的转发数据包,如果某一调频机制在节约能耗的同时,急剧增加了转发的时延,则这种节能机制不能应用于实际系统.所以在对 FASS 的性能测试中,主要关注两点:节能效果和增加的系统延迟.

功耗与频率之间的关系是:$Power=\alpha V^2 f$,其中 $\alpha$ 是依赖于实际器件有关的参数,可以认为是常量,$V$ 是供电电压,$f$ 是工作频率[5].由于 FASS 仅动态调节模块的工作频率,可以简化功耗与频率之间的关系为 $Power=Bf$,其中 $B$ 是与器件相关的参数.因此,我们采用以下的式(5)~(7)计算一个模块的功耗.

$$F_i=\frac{F_{\max}}{2^{k-1-i}} \tag{5}$$

$$P(F_i)=\frac{Count(F_i)}{\sum_{i=1}^{k}Count(F_i)} \tag{6}$$

$$Power=\sum_{i=1}^{k}\frac{P(F_i)}{2^{k-1-i}} \tag{7}$$

其中,$F_{\max}$是模块的最高工作频率;$k$ 是调节的频率档数;$F_i$ 表示第 $i$ 档的频率;$Count(F_i)$是对第 $i$ 档频率的计数值;$P(F_i)$是归一化后,模块工作在第 $i$ 档频率的时间比例.

在上一节中,我们已经详细介绍了实验中所使用的 NetFPGA 平台.特别地,我们将 NetFPGA 的核心时钟频率配置为其默认值 125 MHz,使得 NetFPGA 最高能有 8 Gbps 的吞吐率.然而在我们的实验中,只利用了 4 个以太接口(并没有使用 CPU-DMA 通路),使得 NetFPGA 实际上能达到的最高吞吐率为 4 Gbps.为了能够衡量节能效果和增加的传输时延,我们在 NetFGPA 上实现了两个 IP 路由器:一个是 NetFPGA 默认的参考路由器,一个是应用了 FASS 机制的路由器.对于 FASS 路由器,我们使用时钟适配器来调整时钟的大小.为了防止低速模式下可能出现的数据溢出,我们在所有实验中将缓存的大小设置为 16 KB.

为从实验角度验证第 3 节中理论分析的结果,对于 FASS 路由器我们采用 4 组不同的配置.在所有的 4 组配置中,$M=5$,且均采用幂为 2 的指数分频方式.应用时钟适配器的模块,其频率可被指数分为 6 档,使得最低能获得 $\frac{125}{2^5}$ MHz 的输出频率,对应于 250 Mbps 的吞吐率.对应于 6 档频率,一共有 5 组双门限阈值,在 4 组配置中分别按照指数和均匀形式进行分布,其中最高阈值 $T_{MH}$ 均为 512(对应于 4 KB).对于高低阈值差,如同第 3 节的理论分析,分为 $A$ 和 $B$ 两组.具体地,4 组配置可总结在表 1 之中(其中 $M=5$,$T_{0H}=0$).

**表 1 实验中测试的不同阈值的设置**

| | $T_{iH}(1\le i\le M)$ | $T_{iL}(1\le i\le M)$ |
|---|---|---|
| Reference | 无 | 无 |
| *A-ERUT* | $(i/M)T_{MH}$ | $T_{iH}-4$ |
| *A-ERET* | $2^{-(M-i)}T_{MH}$ | $T_{iH}-4$ |
| *B-ERUT* | $(i/M)T_{MH}$ | $T_{(i-1)H}+4$ |
| *B-ERET* | $2^{-(M-i)}T_{MH}$ | $T_{(i-1)H}+4$ |

在给出如上 4 组配置后,我们使用一台思博伦 SmartBits 600 网络性能分析仪来生成实验流量并分析系统性能.在我们的实验中,NetFPGA 的 4 个 1GE 端口都被使用.输入流量从最大输入流量的 5%调节至 95%,步进为 5%,特定流量下都采用随机数据包发送方式.对于每一个不同大小的流量,我们进行 30 s 的测试,记录系统的平均时延和应用

FASS 的模块的频率分布. 为了使得实验数据可靠，所有的实验均重复 10 次以获得平均结果.

图 8 中展示了参考路由器和 FASS 路由器在 4 个不同配置下，所处于不同流量输入的平均响应时间(包括队列时延和例行的处理时间). 由图 8 所示，FASS 路由器 4 个配置下的时延曲线与理论模型的曲线一致. 其中，在所有的输入情况下，指数频率指数阈值的配置，其时延都低于指数频率均匀阈值的配置，这验证了阈值的分布应与频率的分布相匹配的推论. 另一方面，$B$ 组的时延均略低于 $A$ 组的时延，这支持了高低阈值差应设置为尽可能大的结论. 在图 8 中，在某些负载下，平均延迟会发生突变，是因为在这种负载下，队列的队长会恰好陷入僵局，即一直运行在低频，但又不能突破队列长度来切换到更高频率，从而使得平均时延加大.

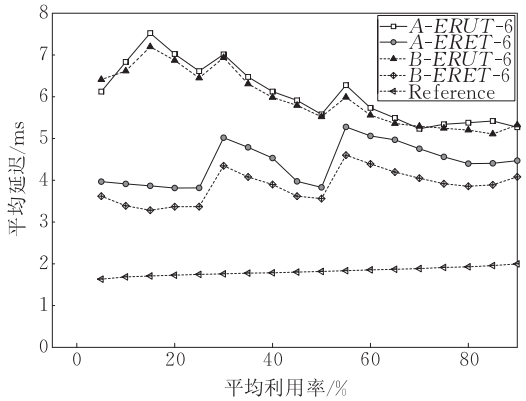



图 8　参考路由器和 FASS 路由器在不同配置、不同负载下的时延(最大负载为 4 Gbps)

然而，在实际实验中的时延要远大于参考路由器的时延，即便是在处理速率更快的情况下，这仍然与理论分析不相吻合. 我们认为这可能由三方面的原因所致. 首先，实际实验中数据包到达可能并不严格服从理论分析中假设的泊松分布；其次，实际频率切换所需的时间并不一定可以忽略，这有待于进一步研究；最后，也是最有可能的，理论分析中对应的结果只适用于一个模块，而在我们的硬件实现当中，有多个模块应用了 FASS 策略中的调频机制.

总体而言，由于双门限多阈值多频调频策略的引入，整个路由器的系统响应时间有所增加，但始终处于一个可接受的范围. 而对于阈值的分布，应与频率的分布相匹配，且同组高低门限的差异应尽可能设置大，以减少路由器进行频率切换的次数，使得系统尽可能多地工作在稳定的状态下.

图 9 中显示了应用 FASS 的模块的动态相对功耗. 由于受限于测量仪器，模块级功耗我们不能直接测得，因而我们根据记录的频率分布按照式(8)推出模块的动态功耗. 从图中可得，在所有的流量输入下，FASS 路由器都能节省大量的动态功耗. 相反地，由于参考路由器始终工作在最高频率，因而即使在输入流量很低的时候，参考路由器依然会消耗非常多的能量. 然而读者可能会产生疑问，为什么当流量负载接近 100% 时，相对的动态功耗只有不到 50%？ 正如在第 4 节所介绍的，NetFPGA 的用户数据通路与 8 个接收/发送队列相连，其中 4 个是我们并没有用到的 CMU-DMA 通路. 由于路由器的核心频率配置为 125 MHz，实际上路由器数据通路支持的吞吐率为 8 Gbps，因而对应的功耗也和流量为 8 Gbps 时相等. 然而由于系统的最大输入流量为 4 Gbps，因而相对的动态功耗最大值也应在 50% 左右. 然而对于参考路由器，即使是在空闲状态下，其所有的模块均工作在最高频率，这导致了与 FASS 路由器相比较大的动态功耗. 另一方面，由于频率只对设备的动态功耗有影响，实际对应的节能效果可能并不如图 9 中所示的那样理想. 但是，通常情况下在不支持频率调节的设备中，其动态功耗远大于静态功耗. 因而，FASS 机制可以使得支持它的模块节省可观的能量，特别是当设备和模块的平均利用率较低的时候.

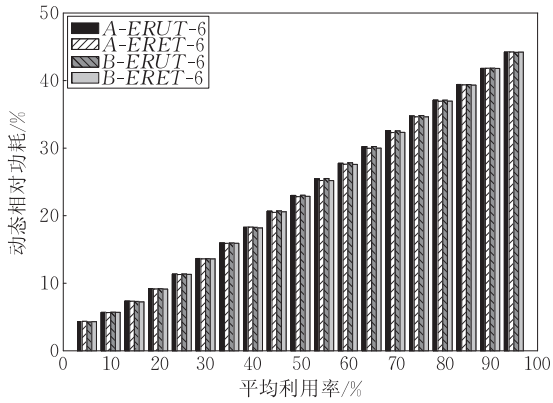



图 9　应用 FASS 的模块在不同负载下的动态相对功耗(最大负载为 4 Gbps)

图 10 中显示了 $B$ 组的指数频率、指数阈值配置在不同流量输入下所对应的平均速率分布，其中柱状图的每一列表示归一化后各个频率运行时间的比例. 可以看出，随着输入流量的不断增大，模块工作在较高频率的时间会有所增加. 而即便当输入流量接近最大负荷时，设备也并不需要时刻工作在对应的处理速率上. 这是因为在实际的网络环境当中，数据包的到达都存在间隙，这一阶段中设备所面临的处理压力相对较小，因而可以工作在相对较低的

频率上,同样可以达到节能效果.而一旦流量的输入速率超过了当前的工作速率,设备则会切换回对应的甚至更高的频率进行处理,以满足性能需要.因此,FASS 确实能使设备的工作速率适应其工作负载,这样可以节约大量能量.对大部分网络设备而言,它们在多数时间都处于较低利用状态,这样通过应用 FASS 机制,就能使模块大部分时间工作在较低频率,从而在维持足够的系统性能的前提下,使这些设备上节约大量能量.图 11 显示了在 1%~30% 负载下,$B$ 组的各个频率的比例.可以看到当流量小于 10%时,90%的时间,模块工作在最低的两档频率;当流量小于 30%时,80%的时间,模块工作在最低的三档频率.

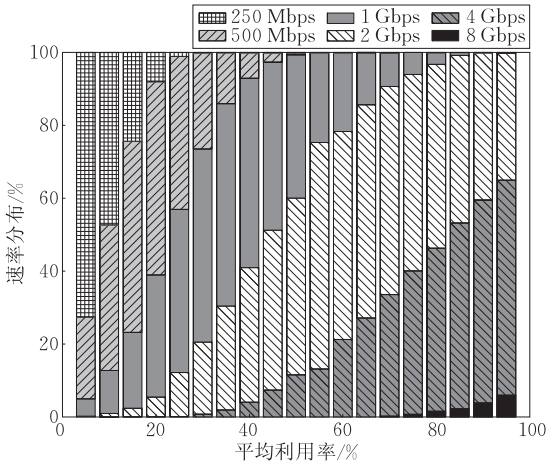



图 10 $B$ 组指数阈值配置在不同负载下的平均速率分布(最大负载为 4 Gbps)

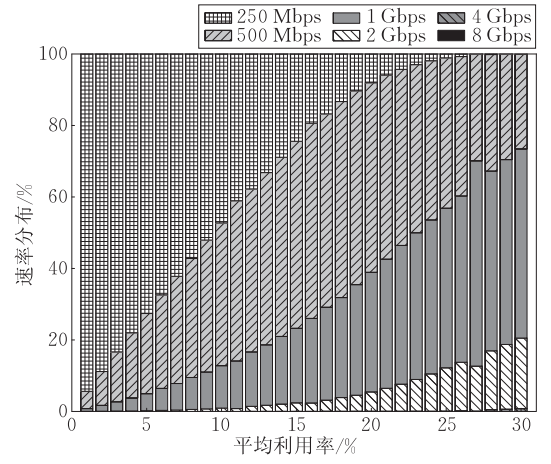



图 11 $B$ 组指数阈值配置在 1%~30%负载下的平均速率分布(最大负载为 4 Gbps)

## 6 相关工作

在过去的几年中,降低网络设备的能耗已经成为网络体系结构设计、路由器和交换机等网络设备设计的重要指标,引起学术界和工业界的广泛关注,有许多工作都围绕节能、绿色互联网展开.Nedevschi 等人[3]提出的两种设备能耗管理方案,被研究人员广泛认可和使用.第 1 种将处于空闲状态的模块休眠,例如路由器控制平面的 CPU、内存模块等;第 2 种根据模块的负载动态地调整模块的工作频率,例如将 CPU 从全速 100%工作调整为半速 50%工作.一般认为休眠方式具有较好的节能效果,而调频方式则更适合负载不断变化的网络环境.而 Wierman 等人[9]在处理器共享系统中研究、应用调频机制,实验结果表明在相同的节能效果下,动态调频机制能够显著地改进系统处理突发流量的鲁棒性.Gupta 和 Singh[2]设计的动态链路关闭算法,能够根据当前系统数据包处理等待队列的长度,结合预测机制,动态地关闭或开启设备的以太网接口,从而实现节能.而根据 Tucker 等人[10]和 Neilson[11]对路由器、交换机中各个物理部件能耗的调查,以太网设备的接口器件仅占整个设备能耗的 13%,所以需要其他技术来实现除接口外其他物理器件的节能.Gunaratne 等人[4]第一次提出采用双阈值来动态调整频率的思路,能够有效解决因为单阈值引起的频率不断切换(抖动)的问题,从而大大降低了系统的整体时延和能耗.但双阈值策略依然存在阈值选取等问题.

NetFPGA[12]提供了一个能够让研究人员快速学习、使用的工业级系统设计、开发平台.它以开源方式提供多个路由器[8]、交换机[7,13]、网卡、发包仪[14]等参考设计,帮助科研人员验证自己的网络优化设计,大大减少了网络研究的任务量.

## 7 结束语

本文提出了一种快速自适应频率调整的机制 FASS 用于减少网络设备的能耗.FASS 采用双门限多阈值实现频率的平稳切换.一方面避免了单阈值带来的频率抖动问题;另一方面,相比双阈值策略,有效降低了模块能耗和时延.同时,我们通过修改 NetFPGA 现有的 4 端口千兆参考路由器,添加 FASS 到数据包处理模块中,验证了 FASS 的有效性;建立马尔可夫模型对阈值的选取进行了理论分析,并使用 Smartbits600 网络测试仪测试了不同负载情况下 FASS 的性能.马尔可夫模型分析结果和实际实验结果表明 FASS 能在仅增加可容忍的时延

情况下，有效减少模块能耗. 负载较重为 3.6 Gbps（原型路由器的 45%负载率）时，可以降低 60.83%的功耗，伴随 1.26 ms 的时延增加. 对于一般负载小于 1.2 Gbps（原型路由器的 15%负载率）时，可以节约 89.6%以上的能耗，伴随小于 2.92 ms 的时延增加.

## 参 考 文 献


[1] Maruti G, Suresh S. Greening of the internet//Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York, USA, 2003: 19-26

[2] Gupta M, Singh S. Dynamic ethernet link shutdown for energy conservation on ethernet links//Proceedings of the IEEE International Conference on Communications. Glasgow, Scotland, 2007: 6156-6161

[3] Nedevschi S, Popa L, Iannaccone G, Ratnasamy S, Wetherall D. Reducing network energy consumption via sleeping and rate-adaptation//Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation. San Francisco, Canada, 2008: 323-336

[4] Gunaratne C, Christensen K, Nordman B, Suen S. Reducing the energy consumption of ethernet with adaptive link rate (ALR). IEEE Transactions on Computer, 2008, 57(4): 448-461

[5] Kaxiras S, Martonosi M. Computer Architecture Techniques for Power-Efficiency. USA: Morgan and Claypool Publishers, 2008

[6] Meng Wei, Wang Yi, Hu Chengchen, He Keqiang, Liu Bin. Greening the internet using multi-frequency scaling schemes//Proceedings of the IEEE International Conference on Advanced Information Networking and Applications. Fukuoka, Japan, 2012: 128-135

[7] Gibb G, Lockwood J, Naous J, Hartke P, McKeown N. NetFPGA: An open platform for teaching how to build gigabit-rate network switches and routers. IEEE Transactions on Education, 2008, 51(3): 364-369

[8] Lockwood J, McKeown N, Watson G, Gibb G, Hartke P, Naous J, Raghuraman R, Luo J. NetFPGA: An open platform for gigabit-rate network switching and routing//Proceedings of the IEEE International Conference on Microelectronic Systems Education. California, USA, 2007: 160 -161

[9] Wierman A, Andrew L, Tang A. Power-aware speed scaling in processor sharing systems//Proceedings of the 28st Annual IEEE International Conference on Computer Communications. Rio de Janeiro, Brazil, 2009: 2007-2015

[10] Tucker R S, Parthiban R, Baliga J, Hinton K, Ayre R W A, Sorin W V. Evolution of WDM optical IP networks: A cost and energy perspective. IEEE Journal of Lightwave Technology, 2009, 27(3): 243-252

[11] Neilson D T. Photonics for switching and routing. IEEE Journal of Selected Topics in Quantum Electronics (JSTQE), 2006, 12(4): 669-678

[12] Watson G, McKeown N, Casado M. NetFPGA: A tool for network research and education//Proceedings of the Workshop on Architecture Research using FPGA Platforms. Austin, USA, 2006: 1-4

[13] McKeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner J. OpenFlow: Enabling innovation in campus networks. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69-74

[14] Covington G A, Gibb G, Lockwood J W, Mckeown N. A packet generator on the NetFPGA platform//Proceedings of the 17th IEEE Symposium on Field Programmable Custom Computing Machines. Washington, USA, 2009: 235-238



**WANG Yi**, born in 1983, Ph. D. candidate. His research interests include router architecture design and implementation, packet forwarding and named data networking.

**MENG Wei**, born in 1990, undergraduate. His research interests include networking system and network security.

**HU Cheng-Chen**, born in 1981, associate professor. His main research interests include computer networking systems, network measurement and monitoring.

**HE Ke-Qiang**, born in 1987, M. S. candidate. His research interest is switch design and implementation.

**LIU Bin**, born in 1964, Ph. D., professor. His main research interests include computer networking systems, router architecture, network processor and named data networking.


**Background**

Internet routers consume a considerable amount of power. The concept of energy-efficient networking has been a hot research topic in the past few years, gaining increasing popularity. Besides the widespread sensitivity to ecological issues,

such interest also stems from economic needs, since both energy costs and electrical requirements of Internet Service Providers' infrastructures around the world show a continuously growing trend. In addition to electricity bills, the large power consumption by network devices also puts a lot of stress on power delivery to and heat removal from router components as well as the hosting facility. Thus it is crucial to save power on routers for the sustainability of the Internet infrastructure.

Speed scaling scheme has high robustness against burst traffic, nevertheless there are only few components (Ethernet PHY, CPU) in network devices supporting it. Most components in the data path of network devices are driven by the same clock crystal oscillator, and operate at full speed regardless of traffic workload. Since all components work synchronously in most current devices, it is extremely difficult to independently adjust the frequency of single component. Fortunately, asynchronous FIFO could be a desirable substitute for synchronous FIFO, in that the two end components of asynchronous FIFO could work at distinct clock frequencies, making it possible to tune frequencies of multiple components respectively. We have done a lot of researches in this area, and part of the work has been presented at IEEE International Conference on Advanced Information Networking and Applications (AINA) 2012.

In this paper, we have proposed a Fast Adaptive Speed Scaling mechanism aiming at energy conservation in network devices. Especially, the mechanism includes a Multi-Dual-Threshold frequency scaling policy and a fast frequency switch approach. The proposed mechanism is further implemented on the NetFPGA platform for validation with moderate modifications on the reference router. Experiment results indicate that the proposed mechanism effectively cuts off the power consumption of the hardware components inside a router with slight increase in average packet delay.

This paper was partially supported by National Natural Science Foundation of China (Nos. 61073171, 60873250), Tsinghua University Initiative Scientific Research Program, and the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20100002110051).