节能以太网的节能策略综述

蒋万春,廖凯琴
(中南大学计算机学院,湖南长沙410083)

摘 要: 节能以太网是解决当前以太网中日益严峻的能耗问题的标准方案.在节能以太网中,节能策略及其参数配置决定了节能以太网设备进入和退出低功耗状态的时机,是影响数据帧延时和节能效果的关键.近年来,国内外开展了大量关于节能以太网的节能策略研究.本文综述了1/10Gbps和40/100Gbps节能以太网的节能策略.首先从策略设计和建模分析的角度总结了1/10Gbps节能以太网的节能策略相关研究.然后,详细描述了40/100Gbps节能以太网的主要节能策略及其核心设计思想.接着,对比和分析了各种节能策略在节能状态选择、节能时长以及状态转换周期上的优缺点.最后,指出了节能策略在网络流量分布、负载状态以及用户延时需求上所面临的机遇与挑战.

关键词: 以太网标准;节能以太网;节能策略;低功耗状态;M/G/1模型;以太网能耗;网络负载
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112(2021)09-1830-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DXZB.20200681

Survey on the Strategies of Energy Efficient Ethernet

JIANG Wan-chun, LIAO Kai-qin
(*School of Computer Science and Engineering, Central South University, Changsha, Hunan* 410083, *China*)

Abstract: The energy efficient ethernet (EEE) is a current standard that addresses the increasingly severe problem of large energy consumption in Ethernet. The strategies of EEE and their parameter configurations determine the time of EEE entering and the exiting low power idle states, which is crucial to both delay and energy consumption. Recent years, a large number of energy-saving strategies of EEE have been studied at home and abroad. This paper surveys the energy-saving strategies of 1/10Gbps and 40/100Gbps EEE. At first, it analyzes the strategies of 1/10Gbps EEE from the perspective of strategy design and modeling analysis. Subsequently, the main ideas of 40/100Gbps EEE strategies are presented. As a step further, the advantages and disadvantages of various EEE strategies in state selection, energy-saving duration, and state transition period are compared and analyzed. Finally, the challenges and opportunities faced by EEE strategies in network traffic distribution, load status, and user latency requirements are discussed.

Key words: ethernet standards; EEE; energy saving strategy; low power idle state; M/G/1 model; energy consumption of ethernet; traffic load

1 引言

近年来,以太网的使用愈加广泛,越来越多的网络设备接入到以太网中.同时,以太网链路带宽也在飞速增长[1~4].链路带宽的增长使得以太网端口功耗不断上升[5,6].然而,以太网链路的利用率仅有5%~30%[7~9],端口在空闲时间内仍旧处于100%耗能的活跃状态,造成了极大的能源浪费.

节能以太网是目前降低以太网端口能耗的主要方式.使用节能以太网的端口可有效减少80%以上的功耗[10],大大减少电力开销[11].2010年,IEEE 802.3az工作组制定并颁布了1/10Gbps节能以太网标准[12].该标准规定,以太网端口在空闲时将关闭部分传输组件,进入低功耗状态.处于该状态下的端口虽然无法传输数据,但端口所消耗的能量仅为活跃状态下的10%.只有当端口被唤醒进入到活跃状态之后才能对数据进行传输.端口进入和退出低功耗状态所必需的时间分别为2.88μs和4.48μs.2014年,IEEE802.3bj工作组又发布了40/100Gbps节能以太网标准[13].该标准中,端口从低

收稿日期:2020-07-09;修回日期:2020-11-18;责任编辑:覃怀银
基金项目:湖南省自然科学基金青年项目(No.2020JJ5768);中南大学创新驱动计划青年项目(No.2020CX033);国家自然科学基金面上项目(No.61972421);中南大学研究生自主探索创新项目(No.2020zzts601)

功耗状态唤醒的时长由4.48μs增长到了5.5μs. 带宽增长而端口恢复到正常工作的时间却变长,这与用户的低延时需求相冲突. 因此,该项标准新增加了快速唤醒状态,以使处于该低功耗状态下的端口能够快速进入到活跃状态. 处于快速唤醒状态的端口同样无法传输数据,其可节省的能量占活跃状态下的70%. 同时,该项标准将原本的低功耗状态重新命名为深度睡眠状态,端口在该状态下消耗的能量只占活跃状态下的10%. 综上,40/100Gbps节能以太网端口具体的状态转换及相应的延时开销如图1所示.

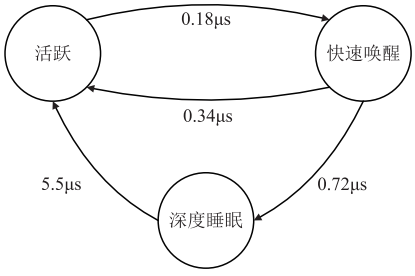


图1　40/100Gbps节能以太网的状态

虽然使用节能以太网标准的端口在低功耗状态下能够节省大量功耗,但在非活跃状态下端口无法对数据进行传输,而恢复到活跃状态又需要一定的时间,这使得数据帧面临额外的延时. 然而,延时增加意味着经济效益降低[14]. 近年来,电子商务,网页搜索,贸易系统等工业互联网对网络的传输延时也提出了更高的要求[15~19]. 因此,利用节能以太网时需要在节能和引入的延时两方面取得折中. IEEE工作组只负责制定指导生产的节能以太网标准,并不负责设计节能以太网使用的策略. 如何设计出性能良好的节能策略,使其在降低能耗的同时不引入较大的延时,成为了当前工业界和学术界共同的研究热点[20~44].

近年来多种针对节能以太网的节能策略被提出. 本文首先将相关研究工作划分为两大类:1/10Gbps节能以太网的节能策略和40/100Gbps节能以太网的节能策略. 本文从策略设计和建模分析的角度详细介绍了1/10Gbps节能以太网相关的节能策略研究工作,并详细地介绍了40/100Gbps节能以太网中基于两种不同低功耗状态所设计的节能策略. 更进一步,本文对比和分析了目前已有的节能策略的优缺点. 最后,本文分析了影响节能策略的关键因素,并指出了节能策略在未来研究发展过程中所面临的机遇和挑战.

2　节能策略的研究现状

目前已经有较多关于节能以太网的节能策略的研究,如图2所示,这些节能策略可被分为两大类:1/10Gbps节能以太网的节能策略以及40/100Gbps节能以太网的节能策略. 我们从自适应负载和节能状态选择等角度对节能策略进行介绍.

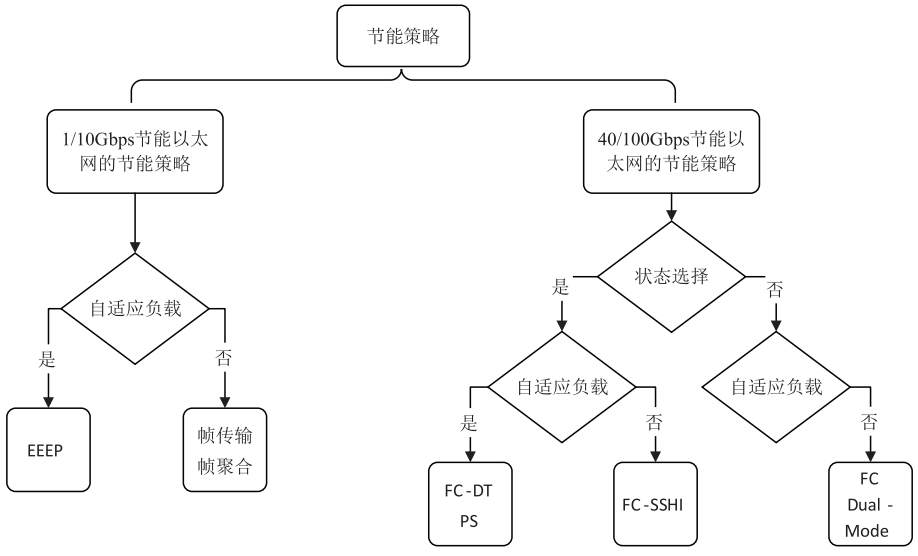


图2　节能以太网中的节能策略分类

2.1　1/10Gbps节能以太网的节能策略

1/10Gbps节能以太网只有一种节能状态,针对其所设计的节能策略需要考虑端口进入和离开低功耗状态的时机. 本节将介绍这些节能策略的设计思想以及围绕这些节能策略展开的建模分析工作.

2.1.1　节能策略设计

目前,1/10Gbps节能以太网中的节能策略主要有3种. 帧传输(frame transmission)策略[20]是其中最直观的方案,其核心思想是端口在空闲时转向低功耗状态,在新的数据帧到达时触发唤醒操作. 在数据传输完毕时,端口

会从活跃状态经过休眠过程过渡到低功耗状态. 端口进入低功耗状态时,会开启一个计时器. 若计时器超时前有数据帧到达,端口会立即被唤醒;相反,若无数据帧到达,端口将在计时器超时后被唤醒,进入到活跃状态.

帧聚合(frame coalescing)策略[21]除了计时器参数外还设置了计数器参数. 在计时器超时或者累积的数据帧数超过计数器阈值后端口才离开低功耗状态.

上述两种策略使用的是静态参数,因而它们都无法适应负载的动态变化. 帧传输策略更注重数据帧的响应时间,只能获得有限的节能效果. 帧聚合策略则增加计数器参数来延长节能以太网端口停留在低功耗状态的时间,进一步提升了节能效果. 不过帧聚合策略下数据帧的延时也进一步增大.

2017年提出的EEEP[22]策略则动态地设定计时器参数以自适应负载的变化. 如图3所示,在每个周期,它通过历史流量的特征分析,使用自回归滑动平均(ARMA)模型等手段预测下一个周期内即将到达的数据帧数,然后计算传输这些帧所需的时间,最后控制节能以太网端口及时退出低功耗状态,以便留出足够的时间传输这些可能到达的数据帧.

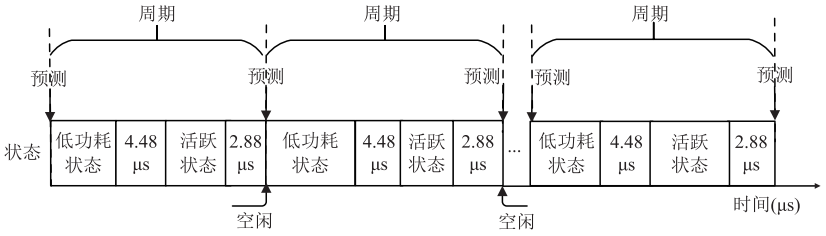


图3 EEEP策略的状态转换过程

周期性的操作使得EEEP策略下的端口在每个周期内停留于低功耗状态下的时间随历史负载的变化而变化. 当然,预测的误差会导致端口的状态转换过程与期望之间存在一定的偏差.

2.1.2 节能策略的建模分析

表1总结了目前节能以太网中的建模分析工作[23~32],它们从理论的角度分析了节能以太网的能耗和延时,为其参数配置提供指导.

表1 节能以太网中节能策略的建模分析

文献序号	年份	分析的节能策略	到达间隔	分析方法	主要分析目标
[23]	2011	帧聚合	泊松分布	M/M/1	能耗
[24]	2011	帧传输	泊松分布	M/G/1	能耗
[25]	2013	帧聚合	泊松分布	M/G/1	平均延时和能耗
[26]	2013	帧聚合	泊松分布	M/G/1	延时约束下的能耗
[27]	2017	帧聚合	泊松分布	M/G/1	计时器和计数器起作用的负载范围
[28]	2016	含双向链路的帧聚合	泊松分布	M/G/1	双向EEE链路下的平均延时和能耗
[29]	2013	只含计时器的帧聚合	泊松分布	M/G/1	延时分布(尾延时)
[30]	2020	帧聚合	复合泊松分布	M/M/1	能耗
[31]	2017	帧聚合(多端口)	叠加的泊松分布	排队网络	网络中的延时分布
[32]	2012	帧聚合	随机	GI/G/1	平均延时和能耗
[37]	2017	FC-SSHI	泊松分布	M/G/1	能耗
[40]	2016	Dual-Mode	泊松分布	M/G/1	能耗
[41]	2016	Dual-Mode	泊松分布	M/G/1	平均延时

假设数据帧的到达间隔服从泊松分布,文献[23]建立了M/M/1模型来刻画帧聚合策略下端口的能耗,而文献[24]则是构建了M/G/1模型来分析帧传输策略下端口的能耗. 虽然它们都是先利用排队论相关知识计算出端口停留在各种状态下的时间占比,再进一步推导得到平均能耗的表达式,但是最终模型呈现的效果不同. 前者能够反映计时器和计数器两个参数在不同的平均负载下对能耗的影响,后者则可以用来衡量帧传输策略下节能以太网端口可实现的节能量.

基于上述假设,节能以太网中也构建了许多针对帧聚合策略的M/G/1模型. 文献[25]提出的M/G/1模型旨在分析计时器和计数器参数的配置对能耗和平均延时两方面的影响. 他们通过建立基于计数器和计时器的马尔科夫模型,计算出了帧聚合策略下的平均能耗和平均延时的表达式. 特别地,当计数器参数为1时,该模型退化为刻画帧传输策略下平均能耗和延时的模型. 而文献[26]先分析数据帧在不同时间内到达端口的概率,再利用全概率公式、拉普拉斯变换及逆变换等

运算得到平均等待时间的表达式. 该文献在模型建立的过程中还考虑了延时约束的影响. 文献[27]则推导了在启动时间内到达节能以太网端口的数据帧数目的分布,再结合广义P-K公式计算出了平均能耗和延时的表达式. 该模型能够用于计算计时器参数和计数器参数起主要作用的负载范围.

同样基于泊松分布的假设,部分研究人员针对几种"特殊"帧聚合策略构建了M/G/1模型. 文献[28]考虑的是双向链路场景下的帧聚合策略的平均延时和能耗. 虽然该文献求解平均能耗表达式的过程与文献[23]类似. 但该文献还通过求解在每种状态或过程中到达的任意数据帧所需等待的平均时间的表达式D_α以及每种状态或过程中到达的平均帧数η_α,得到了双向节能以太网链路中的数据帧的平均等待时间. 对于只含有计时器的帧聚合策略,文献[29]构建了求解延时分布的M/G/1模型,可用于获取尾延时.

考虑到流量的到达间隔并不总是服从简单的泊松分布,文献[30]尝试使用复合泊松分布来描述端口中聚集的字节数,并构建了求解帧聚合策略下端口能耗的M/M/1模型. 另外,当多个以太网端口连接到单个节能以太网端口时,到达节能以太网端口的数据间隔服从叠加的泊松分布. 针对这种情形,文献[31]利用Palm-Khintchine定理来求解节能以太网端口的平均延时表达式. 其建模分析结果显示互联网络中额外产生的延时开销主要受端口采用的节能策略的参数的影响. 文献[32]则为帧聚合策略下的节能以太网构建了更为通用的GI/G/1模型. 该文献使用概率论以及排队论的相关知识求解了节能以太网端口的平均能耗和平均延时的表达式. 该模型能够准确地刻画流量到达间隔随机的情形.

2.1.3　节能策略应用场景的相关研究

在实际应用中通常需要根据具体的场景来调整节能策略及其参数配置以达到更好的性能.

考虑高性能计算对延时的敏感性问题,文献[33]改进了帧聚合策略. 该机制让节能以太网端口在数据传输完成后继续在活跃状态停留一段时间. 如果这段时间内仍然没有数据需要传输,此时链路才进入到低功耗状态. 显然,该机制可以降低状态转换频率,缓解高性能计算集群中的突发流量对节能策略性能的影响,避免了相应的能耗及延时开销.

此外,文献[34]则研究帧聚合策略的计时器和计数器参数在不同MapReduce集群负载下的作业完成时间和节能量表现,并据此给出MapReduce集群中选择节能策略的建议,即1Gbps以太网中的MapReduce集群使用"传统的"节能以太网策略来实现节能,即在数据传输完成时端口立即切换到低功耗状态,有数据帧到达时就离开低功耗状态. 而1/10Gbps以太网中的集群则宜使用帧聚合策略来节省能耗.

2.2　40/100Gbps节能以太网的节能策略

由于40/100Gbps节能以太网定义了两种低功耗状态,因此节能策略在设计时不仅需要考虑何时进入或者退出低功耗状态,而且需要考虑选择哪一种低功耗状态. 目前,为40/100Gbps节能以太网设计的节能策略主要有Dual-Mode[35], FC[36], FC-SSHI[37], FC-DT[38]和PS[39]这5种.

Dual-Mode策略下的端口会在数据传输完毕后进入到快速唤醒状态,并开启计时器T_F. 若T_F内无数据帧到达,端口将进入到深度睡眠状态. 深度睡眠状态下一旦有帧到达,端口立即被唤醒. 相反,若T_F内有数据帧到达,Dual-Mode策略下的端口就退出低功耗状态,开始下一轮数据传输. 而FC策略则通过增加两个计数器参数延长低功耗状态下的时长. 在T_F超时并且数据帧数小于阈值C_F时,端口进入深度睡眠状态. 当数据帧数超过阈值C_D时,端口才退出深度睡眠状态,如图4所示.

Dual-Mode、FC策略都不具备状态选择的功能. 文献[40,41]的建模分析表明Dual-Mode策略下端口的延时较低,但节省的能耗有限. 而FC策略虽然增加了节能的时长,但低负载下的数据帧有很高的延时. 由于T_F、C_F和C_D是静态参数,因此,Dual-Mode、FC策略无法适应网络负载的变化.

FC-SSHI策略考虑端口在非活跃状态的这段时间内到达的数据帧数量是否超过阈值来选择恰当的节能状态. 若该段时间内到达的帧数目大于$C_D/2$,那么端口

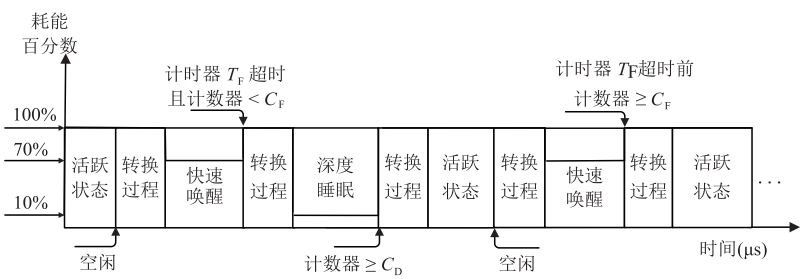


图4　40/100Gbps节能以太网中的FC策略

下次将进入快速唤醒状态，相反，端口将直接进入深度睡眠状态. 具体的状态转换如图5所示.

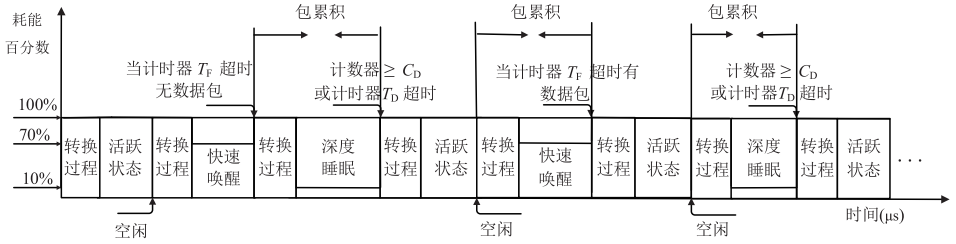


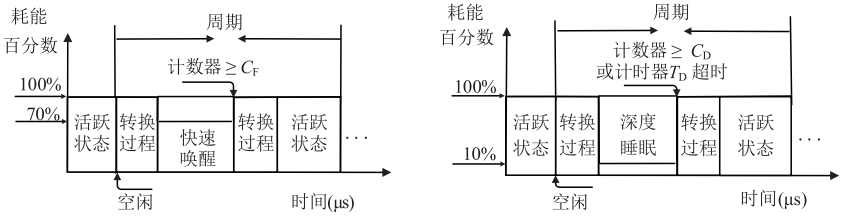
图5 40/100Gbps节能以太网中的FC-SSHI策略

文献[37]构建M/G/1模型分析FC-SSHI策略的平均能耗. 该分析表明，状态选择一定程度上缓解了能耗升高与延时降低之间的冲突. 不过，参数C_D是基于经验设定，因而对低功耗状态的选取存在一定的主观性. 此外，由于它所有的参数都是静态的，无法根据负载或者网络流量的变化做出相应的调整，因而它也无法适应负载的动态变化.

FC-DT策略和PS策略同样具备状态选择的功能. FC-DT策略是根据用户期望的平均延时W^*来选择低功耗状态. 它以文献[31]中的能耗模型和平均延时模型为基础，结合40/100Gbps节能以太网的特性，推导出不同带宽下的目标延时W. $W^*<W$时，FC-DT策略选择快速唤醒状态以确保较低的延时. 相反，策略只选择深度睡眠状态，以节省更多的能量，具体见图6. 而PS策略则根据用户尾延时约束设置期望的周期长度(Expected Time of Cycle, ETC)，每个周期结束时预测下个周期内即将到达的数据帧数，据此计算出活跃状态时长τ，再根据ETC-τ分别计算出两种低功耗状态所对应的能耗从而选择更优的低功耗状态. 具体转换过程见图7. 虽然两者都具有状态选择功能，但FC-DT策略一旦根据用户的期望延时选定了低功耗状态，该策略便只使用一种低功耗状态，具有一定的局限性. 而PS在每个周期内都会选择合适的低功耗状态，这缓解了FC-DT策略在一种带宽下只使用同一种低功耗状态的局限性.

同样，FC-DT策略和PS策略都能够适应负载的变化. FC-DT策略选定低功耗状态后，会根据历史周期内的数据帧数和时长计算出数据帧的平均到达速率，并根据该速率动态地调整计数器参数C_F或C_D，使得节能以太网端口能够根据流量负载信息选择适当的时机退出低功耗状态，从而在确保延时不超过W^*的基础上尽可能地节省能耗. 而PS策略的负载自适应性则受益于它的预测机制. 它使用ARMA模型或EWMA方法预测下个周期内需要的数据传输时间，并动态地调整端口处于低功耗状态下的时间. 即PS通过动态地调整每个周期内的计时器参数来适应负载的变化.

综上，5种节能策略的要点总结见表2.



(*a*) 用户期望的平均延时小于目标延时 (*b*) 用户期望的平均延时大于目标延时

图6 40/100Gbps节能以太网中的FC-DT策略

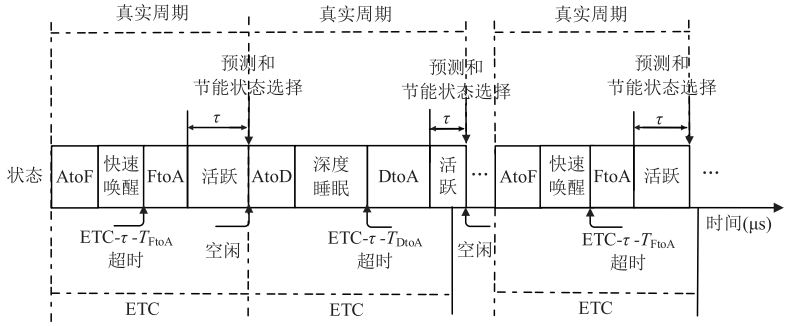


图7 40/100Gbps节能以太网中的PS策略

表2　40/100Gbps节能以太网的主要节能策略

策略	年份	核心思想	主要参数
Dual-Mode[35]	2015	无帧传输时转入快速唤醒状态,启动计时器,在计时器超时前如果仍然无帧到达,进入深度睡眠状态	计时器T_F
FC[36]	2015	在Dual-Mode策略的基础上,聚合快速唤醒和深度睡眠等低功耗状态下到达的数据帧,延长停留在低功耗状态的时长.	计时器T_F 计数器C_F,C_D
FC-SSHI[37]	2017	在FC策略的基础上,根据历史流量信息动态地选择一种低功耗状态.	计时器T_F,T_D 计数器C_D
FC-DT[38]	2017	在FC策略的基础上,根据历史流量信息动态地配置参数C_F和C_D,根据用户期望的平均延时选择节能状态.	计时器T_D 计数器C_F,C_D
PS[39]	2020	参考用户对尾延时的约束设置期望周期长度,根据历史周期的流量信息进行状态选择以及动态控制端口在节能状态下时长.	期望的周期长度ETC

3　节能策略的对比和分析

3.1　节能状态选择

40/100Gbps节能以太网标准定义了两种低功耗状态,基于该标准的节能策略面临节能状态选择的问题. 如表3所示,Dual-Mode策略和FC策略将快速唤醒状态作为深度睡眠状态的前置状态,再根据数据帧的到达情况决定端口是进入深度睡眠状态还是退出快速唤醒状态. Dual-Mode策略在T_F时间内无数据帧到达时进入到深度睡眠状态,FC策略在T_F时间内到达的数据帧小于C_F时进入到深度睡眠状态. 与之不同的是,FC-SSHI策略的节能状态选择并非参照T_F时间内到达的数据帧数,而是通过观察上一个周期内到达的数据帧数是否超过$C_D/2$来进行选择. 若据此选择了快速唤醒状态,但T_F时间内到达的数据帧数目小于预期,再像FC策略一样转向深度睡眠状态. 另一方面,FC-DT策略则根据用户对平均延时的需求来选择具体使用的节能状态. PS的节能状态选择不仅考虑了用户的尾延时约束,同时考虑了历史周期中到达端口的数据量. 总之,结合流量模式和用户的延时需求有助于做出更好的节能状态选择策略.

3.2　停留在节能状态的时长

节能以太网停留在节能状态的时长是影响节能策略的节能量以及延时开销的关键因素. 停留时间越长,节能效果越好,但是延时开销相对越大. 表3对不同节能策略在节能状态停留的时长范围进行了总结.

表3　节能策略的对比和分析

策略	节能状态选择	节能状态下停留时长	状态转换周期
帧传输	LPI	$[0, T_F]$	$[T_{trans}, T_F+T_{trans}+\tau]$
帧聚合	LPI	$[0, T_F]$	$[T_{trans}, T_F+T_{trans}+\tau]$
EEEP	LPI	$[0, T]$	$[T_{trans}, \infty]$
Dual-Mode	先进入快速唤醒再到深度睡眠	$[0, \infty]$	$[T_{trans}, \infty]$
FC	先进入快速唤醒再到深度睡眠	$[0, \infty]$	$[T_{trans}, \infty]$
FC-SSHI	快速唤醒或深度睡眠	$[T_F, T_F+T_D]$	$[T_{trans}+T_F, T_{trans}+T_F+T_D]$
FC-DT	快速唤醒或深度睡眠	C_F溢出所需时长或$[0, T_D]$	$[T_{trans}, \infty]$或$[T_{trans}, T_{trans}+T_D]$
PS	快速唤醒或深度睡眠	[0, ETC]	预测准确度越高越接近ETC

帧传输策略通过设置一个静态的计时器T_F来控制端口停留在节能状态的最大时长. 若在计时器消耗完毕时都无数据帧到达,停留在节能状态的时间长度达到最大值. 否则,该策略下的节能以太网端口停留在节能状态的时间长度是不固定的,主要由随机到达的数据帧决定. 因此,帧传输策略停留在节能状态下的时长在$[0, T_F]$区间内.

帧聚合策略通过累积数据帧的方式来延长停留在节能状态下的时长. 具体的停留时长由计时器和计数器参数共同决定. 在高负载时,停留时长主要由计数器阈值决定,且随负载变化而变化,不过上界仍为T_F. 低负载时,停留时长主要由计时器参数阈值决定,是一个固定值. 显然,计时器阈值是停留在节能状态的时长的最大值. 综上,帧聚合策略下端口停留在低功耗状态下的时长在区间$[0, T_F]$内. 值得注意的是,当计数器参数设值为1时,帧聚合策略退化成帧传输策略,如图8所示.

EEEP策略下端口停留在节能状态的时长由预测算法的结果决定. 令EEEP策略的周期长度为T. 由于当前预测结果受之前周期内的负载影响,所以端口停留在节能状态的时长随负载变化而变化. 不过该时长仍受周期长度的限制,取值范围在$[0,T]$内.

在40/100Gbps节能以太网中,Dual-Mode策略类似

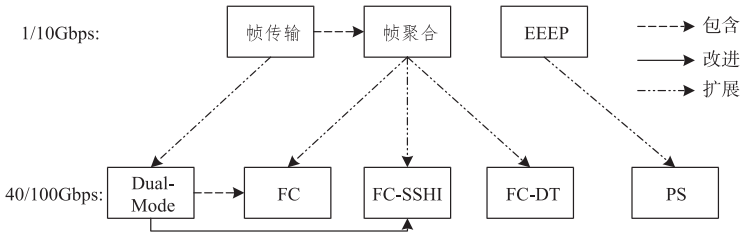


图8 节能策略关系图

于帧传输策略. 若无深度睡眠状态,Dual-Mode策略即退化成帧传输策略. 不过,帧传输策略在节能状态的最大时长受到了计时器参数T_F的限制,而Dual-Mode策略下端口停留在深度睡眠状态的时长是无限制的. 也就是说,若无数据帧到达,Dual-Mode策略下端口将长期停留在深度睡眠状态.

FC策略在参数$C_F=1$并且$C_D=1$时将退化成Dual-Mode策略,如图10所示. 在高负载下,FC策略的计数器C_F的阈值成为影响端口停留在快速唤醒状态的时长的主要因素. 在低负载下,端口更容易转向深度睡眠状态,停留在深度睡眠状态的时长由计数器C_D的阈值决定. 显然,此时节能状态下的停留时长没有上界.

FC-SSHI策略继承了帧聚合策略的思想,它在深度睡眠状态下的时长由C_D和T_D共同决定,T_D是上界. 选择快速唤醒状态时,FC-SSHI策略与Dual-Mode策略类似,它在节能状态下的时长主要由T_F内是否有数据帧到达决定. 综上,FC-SSHI策略下端口停留在节能状态的时长在区间$[T_F, T_F+T_D]$内.

FC-DT策略根据用户的平均延时需求选定节能状态后,状态转换过程与帧聚合策略一样. 不同的是FC-DT策略的计数器参数C_D和C_F根据上一个周期内到达的数据帧数目动态变化,而不是固定值. 另外,如果FC-DT策略根据用户的平均延时需求选择了快速唤醒状态,因为此时没有计时器T_F,端口有可能在快速唤醒状态停留很长时间. 如果选择了深度睡眠状态,端口停留在节能状态的时长上限由计时器参数T_D决定.

在根据用户的尾延时约束配置期望的周期长度后,单个周期内PS的状态转换过程也与帧传输策略一样. 不同的是,每个周期内的计时器大小不一定相同,与历史周期内到达的数据帧数目相关. 当根据历史周期信息预测当前周期内不会有数据帧到达时,计时器具有最大值,即ETC. 而当到达的数据量很大时,计时器往往被设置地很小,以使PS策略下的端口不进入到节能状态中. 如图8所示,当PS策略中只限用一种低功耗状态时,PS策略的状态转换过程与EEEP策略相差无几,不同之处在于EEEP策略的周期长度是凭经验设置的,而PS策略则考虑了用户对尾延时的约束. 综上,PS策略下的节能以太网端口停留在节能状态下的时长在区间[0,ETC]内.

3.3 状态转换周期

节能以太网周期性地工作,若将端口相邻两次离开活跃状态间发生的状态转换过程视为一个周期,则不同节能策略下的周期长度是不同的,如表3所示. 显然,停留在节能状态的时长与周期长度紧密相关.

帧传输策略下周期长度的变化由T_F时间内是否有数据帧到达决定,上界为T_F、转换过程所需时间和活跃状态时长之和. 令T_{trans}表示转换过程所需时间,τ表示活跃状态时长,则帧传输策略下周期长度处在范围$[T_{trans}, T_F+T_{trans}+\tau]$内. 帧聚合策略下周期长度则由参数$T_F$、$C_F$和流量共同决定,上界同样为$T_F$、转换过程所需时间和活跃状态时长之和.

其周期长度的范围仍是$[T_{trans}, T_F+T_{trans}+\tau]$. 不同的是,帧聚合策略下端口在进入活跃状态前缓存的数据帧数目比帧传输策略更多,这将使端口在活跃状态下的时长增大. 帧聚合策略的周期长度根据负载的变化而浮动,而EEEP策略的周期时长为可配置的期望值. 通过动态改变节能状态下的时长来维持相对固定的周期长度. 然而过大的预测误差和突发程度高的流量会使得EEEP对下个周期内的数据传输时间的预测不准确,因而,实际的周期长度与配置的周期长度之间存在着一定的误差,范围为$[T_{trans}, \infty]$.

40/100Gbps节能以太网中,Dual-Mode策略、FC策略、FC-SSHI策略和FC-DT策略测量的周期时长都由计时器、计数器参数和流量共同决定,且动态变化. 由3.2节的分析可得,Dual-Mode策略和FC策略的周期长度范围都为$[T_{trans}, \infty]$. FC-SSHI策略的周期长度范围为$[T_{trans}+T_F, T_{trans}+T_F+T_D]$. 值得注意的是,若流量的到达是均匀的,那么在FC-DT策略下相邻周期的时间长短相差无几. 根据图6可知,当FC-DT策略使用快速唤醒状态时,其周期长度的范围为$[T_{trans}, \infty]$;当它只使用深度睡眠状态时,其周期长度则为$[T_{trans}, T_{trans}+T_D]$. 而PS策略的周期长度则与预测算法的精确度有关. 预测准确度越高,真实周期长度越接近ETC.

4 机遇与挑战

综合分析上述相关研究工作可以发现,节能策略及其参数配置是影响节能以太网端口的节能效果和延时开销的关键. 例如,当前主流的帧聚合相关策略中,

计时器和计数器的值越大,节能以太网端口停留在低功耗状态的时间越长,节能效果越好,但是延时开销也越大. 此外,影响节能策略及其参数配置的主要因素是网络流量分布、负载状态和用户的延时需求. 如图9所示,网络流量分布的差异和负载变动对节能策略及其参数配置的影响非常直接. 当流量负载越重时,帧聚合相关策略中的计时器和计数器应配置越大的数值,以便处于低功耗状态节能的时间长于状态转换的时间开销;相反,流量负载越轻,计时器和计数器应配置越小的数值,以免在低功耗状态进行帧聚合的时间太长,延时开销太大. 因此,针对不同的流量分布和负载状态应当设计不同的参数配置方案以节省更多的能量. 此外,用户的延时需求也影响节能策略设计及其参数配置. 延时要求越严格,意味着节能以太网端口不能在低功耗状态下停留过长的时间,以避免造成过大的延时;相反,节能策略在满足用户延时要求的同时应当尽可能地停留在低功耗状态下,从而实现更好的节能. 然而,网络流量分布和负载状态的影响和用户的延时需求可能出现冲突. 当流量负载很轻的同时用户也需要极低的延时,节能以太网的策略将难以在节能量和延时之间取得恰当的折中.

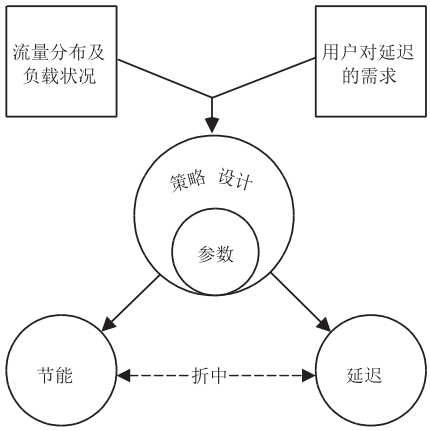


图9　影响节能策略的关键因素

总之,节能策略的节能效果以及延时开销主要依赖于策略的参数配置,间接地受到流量分布和负载状况,以及用户对延时的需求的影响. 综合上述分析和讨论,我们认为未来节能以太网的节能策略研究将面临以下机遇和挑战.

(1) 流量分布或者负载状态已知的特定场景中的节能策略研究. 此时,可以根据已知的网络流量或者负载信息制定简单而高效的节能策略及参数配置方案. 例如,根据高性能计算集群中流量突发的特性,在帧聚合策略的基础上为活跃状态结束后增设一段"缓解"时间,用以降低状态转换频率,从而提升节能效果. 如果用户的延时需求未知,则在策略设计和参数配置时需进一步考虑用户的延时需求对节能策略的间接影响.

(2) 面向用户的延时需求的节能策略研究. 当前的节能策略研究主要关注于如何在不同的网络流量分布和负载状态中取得较好的节能量和延时开销的折中,直接将用户的延时需求引入节能策略设计的相关研究成果较少. 然而,在搜索、实时交互等实际场景中,应用必须满足显式给定的延时需求,以便提升用户体验. 显然,松弛的延时需求将给节能策略设计带来更大的节能空间;相反,严苛的延时约束可能要求节能策略更加温和地追求节能效果. 因此,将用户的延时需求显式地引入节能策略的设计过程,确保满足用户延时需求的前提下最大化节能效果成为研究的目标. 目前,少量相关研究工作将用户的延时需求作为节能策略参数配置的依据,初步地达到这一目标[38,39]. 但是,这些方案无法确保在不同网络流量和负载状态下达到既满足用户的延时需求,又最大化节能效果的目标. 节能策略的设计仍然存在进一步优化的空间.

(3) 面向未知网络流量分布和负载状态的通用节能策略研究. 虽然动态配置计时器和计数器参数的帧聚合策略[25,27],以及根据对未来网络流量的预测结果动态地调整参数的节能策略[37~39]能够在一定程度上适应不同的网络环境,但是帧聚合策略的参数配置和预测算法的应用是建立在网络流量服从泊松分布的基础上. 然而,实际网络环境中流量分布和负载状态未知,使得现有的节能策略的通用性仍然存在进一步的改进空间. 另一方面,随着深度学习等技术的发展,如何在无法事先获知节能策略的应用场景的前提下,自动地抓取流量分布特征以及负载变化,设计出更好的节能策略,成为了值得研究的重要问题.

5　总结

本文论述了当前节能以太网中的相关研究工作,将它们分为1/10Gbps节能以太网的节能策略和40/100Gbps节能以太网的节能策略两大类,从自适应负载和节能状态选择的角度介绍了这两类节能策略,并归纳了针对这些策略所建立的模型. 在此基础上,对比和分析了各种节能策略在节能状态选择、节能时长以及状态转换周期上的优缺点. 最后,本文分析了负载状态、网络环境、用户的延时需求等对节能策略及其参数配置的影响,展望了节能策略研究的发展趋势及其所面临的机遇和挑战.

参考文献

[1] WINZER P. Beyond 100G ethernet [J]. IEEE Communications Magazine, 2010, 48(7):26 – 30.

[2] IEEE P802.3ba. 40Gb/s and 100Gb/s Ethernet Task Force

[EB/OL]. http://www.ieee802.org/3/ba/index.html, 2010.
[3] 张小丹,程丹,徐晶,等. 40G/100G以太网关键技术的研究与应用[J].光通信技术, 2011,35(4):1 – 4.
ZHANG X D, CHENG D, XU J, et al. The research and application of 40G/100G Ethernet key technology[J]. Optical Communication Technology, 2011, 35(4): 1 – 4.(in Chinese)
[4] TORRES-FERRERA P, FERNÁNDEZ-SEGURA O, GUTIERREZ-CASTREJON R. Comparison of 10x40Gbps and 8x50Gbps WDM system for next-generation ethernet operating at 400Gbps[A]. OSA Latin America Optics&Photonics Conference[C]. Cancun Mexico: OSA, 2014.
[5] REVIRIEGO P, CHRISTENSEN K, RABANI LLO J, et al. An initial evaluation of energy efficient ethernet [J]. IEEE Communications Letters, 2011,15(5):578 – 580.
[6] KOHL B. 10GBASE-T Power Budget Summary [R]. Orlando, FL, USA: IEEE, 2007.
[7] BENSON T, ANAND A, AKELLA A, et al. Understanding data center traffic characteristics [A].ACM Workshop on Research on Enterprise Networking [C]. Barcelona, Spain: ACM, 2009.
[8] GUPTA M. A feasibility study for power management in LAN switches[A]. IEEE International Conference on Network Protocols[C]. California, USA: ACM, 2004.
[9] NORDMAN B. EEE Savings Estimates [EB/OL]. http://grouper.ieee.org/groups/802/3/eee_study/public/may07/nordman_2_0507.pdf, 2007.
[10] KHOURY J EL, RONDEAU E, JP GEORGES, KOR AL. Assessing the Impact of EEE Standard on Energy Consumed by Commercial Grade Network Switches[M]. Nancy, France: Green IT Engineering, 2019.
[11] CHRISTENSEN K, REVIRIEGO P, NORDMAN B, et al. IEEE 802.3az: the road to energy efficient ethernet [J]. IEEE Communications Magazine, 2010, 48(11): 50 – 56.
[12] IEEE 802.3az. Amendment 5: Media Access Control Parameters, Physical Layers, and Management Parameters for Energy-Efficient Ethernet[M].New York, USA: IEEE, 2010.
[13] IEEE 802.3bj. Amendment 2: Physical Layer Specifications and Management Parameters for 100Gb/s Operation over Backplanes and Copper Cables[M].New York, USA: IEEE, 2014.
[14] HOFF T. Latency is Everywhere and It Costs You Sales How to Crush It [EB/OL]. http://highscalability. com/Latency-everywhere-and-it-costs-you-sales-how-crush-it, 2019.
[15] DEAN J, BARROSO L A. The tail at scale [J]. Communications of the ACM, 2013, 56(2): 74 – 80.
[16] RUMBLE S M, ONGARO D, STUTSMAN R, et al. It's time for low latency [A]. Proceedings of the 13th USENIX Conference on Hot Topics in Operating Systems [C]. Berkeley, CA, USA: USENIX Association, 2011. 11 – 11.
[17] WILSON C, BALLANI H, KARAGIANNIS T, et al. Better never than late: meeting deadlines in datacenter networks [J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4):50 – 61.
[18] ALIZADEH M, GREENBERG A, MALTZ D A, et al. Data center TCP (DCTCP) [J]. ACM SIGCOMM Computer Communication Review, 2010, 40(4):63 – 74.
[19] LIANG Q, MODIANO E H. Coflow scheduling in input–queued switches: optimal delay scaling and algorithms [A]. IEEE INFOCOM 2017—IEEE Conference on Computer Communications[C]. USA: IEEE, 2017.
[20] REVIRIEGO P, HERNANDEZ J A, LARRABEITI D, et al. Performance evaluation of energy efficient ethernet [J]. IEEE Communications Letters, 2009, 13(9):697 – 699.
[21] REVIRIEGO P, HERNADEZ J A, LARRABEITI D, et al. Burst transmission for energy efficient ethernet [J]. IEEE Internet Computing, 2010, 14(4): 50 – 57.
[22] CENEDESE A, TRAMARIN F, VITTURI S. An energy efficient ethernet strategy based on traffic prediction and shaping [J]. IEEE Transactions on Communications, 2017, 65(1): 270 – 282.
[23] HERRERIA-ALONSO S, RODRIGUEZ-PEREZ M, FERNANDEZ-VEIGA M, et al. A power saving model for burst transmission in energy efficient ethernet [J]. IEEE Communications Letters, 2011, 15(5):584 – 586.
[24] MARSAN M A, ANTA A F, MANCUSO V, et al. A simple analytical model for energy efficient ethernet [J]. IEEE Communications Letters, 2011, 15(7):773 – 775.
[25] MENG J, REN F, JIANG W, et al. Modeling and understanding burst transmission algorithms for energy efficient ethernet[A]. IEEE/ACM 21st International Symposium on Quality of Service (IWQoS) [C]. Montreal, QC, Canada: IEEE, 2013. 1 – 10.
[26] KIM K J, JIN S, TIAN N, et al. Mathematical analysis of burst transmission scheme for IEEE 802.3az energy efficient ethernet [J]. Performance Evaluation, 2013, 70(5): 350 – 363.
[27] PAN XIAODAN, YE TONG, TONY TLEE, et al. Power efficiency and delay tradeoff of 10GBase-T energy effi-

cient ethernet protocol [J]. IEEE/ACM Transactions on Networking, 2017,25(5):2773 – 2787.
[28] CHATZIPAPAS A, MANCUSO V. An M/G/1 model for gigabit energy efficient ethernet links with coalescing and real-trace-based evaluation [J]. IEEE/ACM Transactions on Networking, 2016, 24(5):2663 – 2675.
[29] AKAR N. Delay analysis of timer-based frame coalescing in energy efficient ethernet [J]. IEEE Communications Letters, 2013, 17(7):1459 – 1462.
[30] MAKSI N, BIELICA M. M/M/1 model of energy efficient ethernet with byte-based coalescing [J]. Annals of Telecommunications, 2020, 75(7/8):291 – 305.
[31] RODRIGUEZ-PEREZ M, HERRERIA-ALONSO S, FERNANDEZ-VEIGA M, et al. Delay properties of energy efficient ethernet networks [J]. IEEE Communications Letters, 2017, 21(10):2194 – 2197.
[32] HERRERIA-ALONSO S, RODRIGUEZ-PEREZ M, NANDEZ-VEIGA MFER, et al. A GI/G/1 model for 10Gb/s energy efficient ethernet links [J]. IEEE Transactions on Communications, 2012, 60(11):3386 – 3395.
[33] SARAVANAN K P, CARPENTER P M, RAMIREZ A, et al. Power/performance evaluation of energy efficient ethernet (EEE) for high performance computing [A]. IEEE International Symposium on Performance Analysis of Systems & Software [C]. USA: IEEE, 2013.205 – 214.
[34] SILVA R F E, CARPENTER P M. Energy efficient ethernet on mapreduce clusters: packet coalescing to improve 10GbE links [J]. IEEE/ACM Transactions on Networking, 2017, 25(5): 2731 – 2742.
[35] MOSTOWFI M. A simulation study of energy efficient ethernet with two modes of low-power operation [J]. IEEE Communications Letters, 2015, 19 (10): 1702 – 1705.
[36] HERRERÍA-ALONSO S, RODRÍGUEZ-PÉREZ M, FERNÁNDEZ-VEIGA M, LÓPEZ-GARCÍA C. Frame Coalescing in Dual-mode EEE [EB/OL]. https://arxiv.org/abs/1510.03694, 2015.
[37] MOSTOWFI M, SHAFIE K. Dual-mode energy efficient ethernet with packet coalescing: analysis and simulation [J]. Sustainable Computing Informatics & Systems, 2018, 18(Jun):149 – 162.
[38] HERRERÍA-ALONSO S, RODRÍGUEZ-PÉREZ M, FERNÁNDEZ-VEIGA M, et al. Optimizing dual-mode EEE interfaces: deep-sleep is healthy [J]. IEEE Transactions on Communications, 2017, 65(8):3374 – 3385.
[39] WANCHUN JIANG, KAIQIN LIAO, YULONG YAN, JIANXIN WANG. PS: periodic strategy for the 40–100Gbps energy efficient ethernet [A]. 49th International Conference on Parallel Processing-ICPP [C]. New York, USA: ICPP, 2020. 1 – 10.
[40] SHAFIE K, MOSTOWFI M. An analytical model for the power consumption of dual-mode EEE [J]. Electronics Letters, 2016, 52(15):1308–1310.
[41] MOSTOWFI M, SHAFIE K. Average packet delay in dual-mode EEE: an analytical model [J]. Electronics Letters, 2016, 52(21): 1759 – 1761
[42] 张国强,林森,刘真,等.高能效互联网传输技术研究[J].通信学报, 2012, 33(05):158 – 168.
ZHANG G Q, LIN S, LIU Z, et al. Energy efficient data transmission on the Internet [J]. Journal on Communications, 2012, 33(5): 158 – 168.(in Chinese)
[43] 赵金洲,叶通,李东,等.节能EPON中一种新的ONU休眠策略[J].光通信技术, 2016, 40(005):8 – 10.
ZHAO J Z, YE T, LI D, et al. Novel sleep scheme for ONUs in energy efficient EPON [J]. Optical Communication Technology, 2016, 40(5): 8 – 10.(in Chinese)
[44] AKSIC M, BJELICA M, Packet coalescing strategies for energy efficient ethernet [J]. Electronics Letters, 2014, 50(7):1759 – 1761.

作者简介

蒋万春　男,1987年生于湖南永州.现为中南大学计算机学院副教授、博士生导师.主要研究方向为网络与分布式计算、云计算、拥塞控制、以太网.
E-mail: jiangwc@csu.edu.cn

廖凯琴　女,1995年生于江西宜春.现为中南大学计算机学院硕士研究生.主要研究方向为超高速节能以太网.
E-mail: kaiqinliao@csu.edu.cn