

**Collaborative Research (NeTS-NBD): Increasing the Energy
Efficiency of the Internet with a Focus on Edge Devices**

Ken Christensen

Department of Computer Science and Engineering
4202 East Fowler Avenue, ENB 118
University of South Florida
Tampa, FL 33620
christen@csee.usf.edu

Alan D. George

Department of Electrical and Computer Engineering
327 Larsen Hall, POB 116200
University of Florida
Gainesville, FL 32611-6200
george@hcs.ufl.edu

A proposal for NSF NeTS (NSF-05-540) – funded from August 2005 to August 2008

Project Summary

This proposal addresses the increasingly critical need to improve the energy efficiency of the Internet by focusing on the primary and often neglected energy consumer, edge devices. Studies by Lawrence Berkeley National Laboratory (LBNL) show that about 74 TWh/yr of electricity (which is approximately \$6 billion per year) is consumed by the Internet in the USA alone, of which 24 TWh/yr or 32% could be saved with full use of power management on desktop computers, currently the most common of edge devices on the Internet. Unfortunately, due to limits of existing protocols and architectures, networked desktop computers typically remain powered-up during frequent and often lengthy periods of idleness. As network devices, they are prevented from operating in an energy-efficient manner due to their need to respond to network transactions of various types without warning. The NSF is currently funding projects to investigate energy-use reduction in first-level LAN switches, server clusters, and supercomputers, but severely lacking is research to reduce energy consumption of the largest portion of Internet energy consumers, the edge devices.

Our *approach* to addressing this challenge is to investigate and exploit a synergistic set of novel research concepts for protocol and subsystem infrastructure, and algorithms for effectively controlling them based on traffic and system constraints, so that desktop and other edge devices can be put to sleep during periods of relative idleness while network connectivity is maintained by a low-power hardware proxy integrated into the system. Moreover, our approach also promises to provide additional increases in energy efficiency by reducing consumption of network-related resources during active periods where graceful degradation of performance is acceptable, in effect trading off speed for energy. In addition to desktop computers, this approach will lead to similar solutions for a wide variety of emerging wired and wireless edge devices such as television set-top boxes, network appliances, remote cameras, etc.

The *novel concepts* in our approach include: (1) protocol proxying in the network interface of a desktop computer and/or within a first-level LAN switch to reduce minor use of system resources of the edge device and thus allow system power management to be fully exploited; (2) “smart” wake-up methods to allow power-managed devices to be awoken transparently and only when needed by existing applications and protocols, along with new power-management notification semantics for future network applications; (3) adaptive link rate in Ethernet to trade-off performance (or QoS) for energy efficiency by dynamically operating links at varying data rates (e.g., 1 or 10 Gb/s only when needed, and otherwise at 10 or 100 Mb/s); and (4) architectures for fixed or reconfigurable levels of staged hardware functionality to realize energy-efficient operation via adaptive spatial or temporal assignment of hardware resources to network transactions with power, performance, and functionality scaling.

Our *methodology* to achieve these energy savings is measurement, analysis, modeling, and prototyping of new software and hardware. We will build and evaluate a prototype of an advanced Ethernet interface for edge devices based on adaptive levels of functionality including proxying, smart wake-up, adaptive link rate, and reconfiguration. We will prototype software applications that use new semantics for power management notification across a network. We will *disseminate our research outcomes* to standards groups, government agencies, and industry. We will propose fast autonegotiation mechanisms for adaptive link rate to the IEEE 802.3. We expect to play a key role in influencing future EPA Energy Star specifications for desktop computers and other edge devices (such as, potentially, television set-top boxes). This project will be conducted in a *collaborative partnership* between the University of South Florida and the University of Florida. Four doctoral students will be supported of which one will be targeted to be from an underrepresented population group.

Intellectual merit: The work proposed is intellectually meritorious in that it is the first significant effort to achieve energy savings for the global Internet in terms of its dominant factor, the network-connected edge devices. This research will define the importance of energy efficiency of network edge devices as an economic and environmental issue and result in the integrated design of power management and network applications, protocols, and architectures.

Broader impact: The broader impact of this work is threefold. First is the impact upon society within a few years of completion of this project, enabling significant reductions in energy costs here in the USA and abroad, and supporting the expansion of the Internet into the developing world by reducing operating costs. **Savings in the hundreds of millions to billions of dollars per year in the USA will be achieved if existing power management capabilities can be enabled in network edge devices based on the new ideas in this proposal.** We have existing relationships to allow us to work closely with LBNL, EPA, and industry to disseminate our research outcomes and achieve the expected energy savings. The second broader impact is on the educational process, both in terms of the graduate students directly involved in this project as well as the many university students that will benefit from the course materials that will be created and shared from this research. Finally, through the NSF REU and RET programs, outreach will be made to K-12 and underrepresented populations.

Project Description

1 Introduction

Herein we propose the first major research effort targeting energy savings in the global Internet by focusing on the dominant factor, edge devices. We investigate the feasibility and impact of several novel and aggressive techniques for improved energy efficiency at the edge including protocol proxying, energy state semantics, and adaptive configuration. We will build an adaptive, energy-efficient NIC to demonstrate the significant research ideas. We will disseminate our results to standards, government, and industry.

A growing expense and impact of the Internet is its energy use. The largest consumers of electricity are edge devices including desktop computers where existing power management capabilities – e.g., as mandated by the EPA Energy Star specification [17] and implemented in nearly all modern desktop computers – are usually disabled due to network considerations. Network connectivity induces wasteful energy use in edge devices [14]. The Environmental Energy Technologies Division of the Lawrence Berkeley National Laboratory (LBNL) describes the energy use situation of IT equipment in 2000 as [35]:

We found that total direct power use by office and network equipment is about 74 TWh per year, which is about 2% of total electricity use in the U.S. When electricity used by telecommunications equipment and electronics manufacturing is included, that figure rises to 3% of all electricity use (Kooimey 2000). More than 70% of the 74 TWh/year is dedicated to office equipment for commercial use. We also found that power management currently saves 23 TWh/year, and complete saturation and proper functioning of power management would achieve additional savings of 17 TWh/year. Furthermore, complete saturation of night shut down for equipment not required to operate at night would reduce power use by an additional 7 TWh/year.

That the energy use of the Internet is only 2% of the total may appear to be a small issue, but IT equipment is the fastest growing energy consumer in office buildings [48]. The energy consumption of the Internet has increased since 2000 and it is continuing to increase [25]. In Germany it is estimated that energy consumption by IT equipment will be between 2% and 5% in 2010 [53]. Currently, 9% of the plug load of a typical office building is for IT equipment (most of this equipment are desktop computers and other network edge devices) [48]. The absolute quantity of energy consumed is large; it is equal to the output of several nuclear power plants (at approximately 7 TWh/year for a nuclear power plant [54]). **If an energy savings of the magnitude described by LBNL can be achieved with little initial investment and no perceivable impact to the IT equipment user, it should be aggressively pursued for both economic and environmental considerations.** The NSF (including CISE) is currently funding research in energy reduction in LAN switches [50], server clusters in data centers [7], and supercomputer installations [8]. Significant research funding has been made to investigate energy efficiency for mobile devices (e.g., ad hoc and sensor networks) and of processors. The expected outcomes from this already funded research are complementary to what we propose, but also very significantly do not solve the problems specific to network edge devices including Ethernet-connected desktop computers. In addition to commercial use of desktop computers, residential networks for always-on communications, media playing, and multi-player gaming present an emerging trend in increasing power consumption with significant impacts to home user electricity costs [9]. For example, “Absent significant efficiency improvements, we predict that national set-top box energy use will surpass 40 TWh/yr by 2010.” [47].

It is clear that the artifacts of computer science and electrical engineering are becoming major consumers of the earth’s resources. The NSF has the potential to fund fundamental and groundbreaking research in power management of existing and future Internet-connected edge devices. Such research – even with only modest success – will save many TWh/yr of electricity. The economic and environmental gains will be enormous. **The attached letters from Bruce Nordman at LBNL and Craig Hershberg at the EPA support this expectation of significant economic and environmental gains from the research in this proposal.**

1.1 Objectives and directions for this research

The *objective* of this research is to investigate new ideas that will measurably reduce the energy consumption of edge devices in the Internet. Our focus is on wired – i.e., Ethernet – devices, but many of our ideas will directly

apply to 802.11 and other wirelessly connected edge devices. We look at NICs for existing desktop computers and LAN switches, and for future network edge devices such as intelligent appliances, television set-top boxes, security cameras, and embedded controllers in both commercial and residential applications. The two *fundamental questions* that we ask are:

- 1) What can be done to increase the use of power management in desktop computers and other edge devices in order to reduce energy consumption during long idle periods?
- 2) What can be done to reduce the inherent energy consumption of high-speed links and NICs as a function of demand during low demand or short idle periods?

Our guiding principal is one of energy consumption as a function of user demand. When user demand is low, energy consumption should also be low. This is currently not the case for many network edge devices – such as desktop computers and Ethernet NICs and links – that consume roughly the same amount of energy whether being actively used or idle. We will investigate how the following methods can reduce energy use by edge devices:

- *Protocol proxying* in a NIC and/or within a first-level LAN switch to reduce the need to access the system resources of a desktop computer or other edge device and thus allow power management to be enabled and used during the now extended idle periods.
- *Smart wake-up methods* to allow power-managed devices to be woken-up transparently from a low-power sleep state by existing applications and protocols and only when needed. New power-management notification and wake-up semantics for future networked devices and applications will also be studied.
- *Adaptive link rate* and the achievable energy savings by operating Ethernet links at high data rates (1 or 10 Gb/s) only when needed and otherwise at 10 or 100 Mb/s.
- *Staged levels of functionality* in a “smart” NIC to support diverse proxy requirements with energy efficiency, based on spatially and/or temporally dynamic assignment of hardware resources, dynamic power management, and voltage-frequency scaling.

Our research direction will be centered on using real traffic traces to explore how proxying, smart wake-up, power state notification, adaptive link rate, and adaptive hardware configuration can reduce energy use. We will develop a prototype Ethernet NIC that embodies our ideas and evaluate it using real network applications. Our contention is that improved energy efficiency of network edge devices is possible without affecting user-perceived performance or QoS including application response time. Improved energy efficiency will benefit the individual user with lower cost of operation and will benefit society by reduction of environmental impacts from electricity generation. By reducing the cost of operation, this research can enable expanded Internet deployment in the developing countries.

1.2 Organization of this proposal

The remainder of this proposal is organized as follows. Section 2 describes research challenges in power management of edge devices. The scope and significance of the proposed research are also described. Section 3 reviews existing work in power management of desktop computers and funded work in improving the energy efficiency of the Internet. Section 4 is the research plan where the research challenges outlined in Section 2 are addressed. Section 5 describes the expected outcomes for the project, dissemination of research results including influence on future EPA Energy Star specifications. We address both intellectual merit and broader impacts to society in this section. Section 6 describes results from prior NSF support for the two PIs.

2 Research Challenges

Power management of a computer is possible whenever it is idle. Idle periods can be defined over multiple time scales. The time scales are CPU and instruction level (nano to microseconds), inter-packet (micro to milliseconds), and inter-flow (seconds to hours) where a flow can be a TCP connection comprising multiple packets. Inter-flow idle times are a conservative estimate of available idle time for power management. Even within flows, there are

times between packet transmit and receive that can be used for power management. Methods have been explored to batch web transactions to increase idle time [22]. **Predicting, controlling, and making the best use of idle times is a key challenge to power management.** For each time scale, different power management methods are possible. At the instruction and hardware levels, clock frequency scaling can be exploited to slow-down completion of tasks (and thus reduce energy consumption). Deadline-driven tasks need to complete at their deadline, but there is no value in an early completion that comes at a cost of unnecessary energy consumption. At the inter-packet level there are opportunities to keep a power-consuming transmitter “off” and/or powering-down parts of a NIC. At the inter-flow time scale it becomes possible to power-down the entire edge device. For all time scales, unnecessary computation and communication should be eliminated to save energy.

Desktop computers are ubiquitous in almost all commercial workplaces and residences. Most office desktop computers do very little useful work at night, but remain fully powered-on [35]. As a preliminary work, we investigated the amount of inter-flow idle time in University of South Florida (USF) dormitory PCs. These PCs contribute the majority of traffic to the university Internet connection and are known to be running modern file-sharing programs such as Kazaa, eDonkey, etc. On March 27, 2003 Cisco NetFlow traces were collected for 24 hours from the USF dormitories, it was found that the dormitory machines sent and received 1350 GBytes, which was about 60% of all USF network traffic for the day. Figures 1 and 2 show the characterization results for bytes sent and received and inter-flow idle time for the top 100 PCs in traffic volume. These results show that there is considerable idle time for power management even in what are likely the busiest desktop computers on campus.

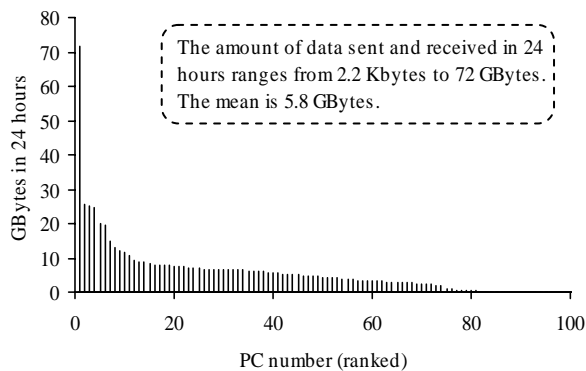


Figure 1. Bandwidth usage of USF PCs

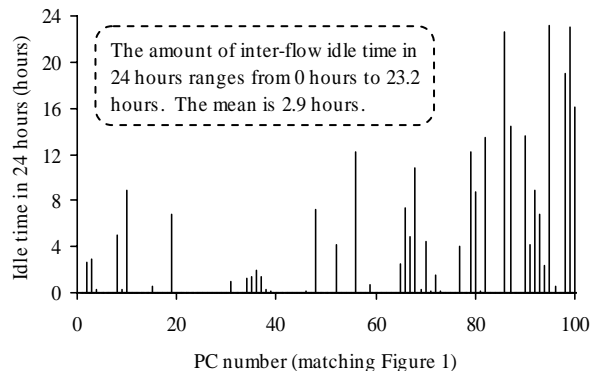


Figure 2. Inter-flow idle time of USF PCs

2.1 Open questions related to the reduction of energy consumption

The initial results based on a characterization of network traffic generated by USF dormitory computers raise several important questions:

- *What are the characteristics of idle periods in network traffic at different time scales?* The research challenge is to collect and characterize representative traffic from existing LAN environments and applications (both commercial and residential). We also need to consider the characteristics of future applications that may drive network edge devices including video to and from television set-top boxes, video from security cameras, and so on. Characterization of idle periods must be done for multiple time scales including inter-flow (e.g., for allowing devices to enter a sleep state and/or adaptive link rate control of an Ethernet link) and inter-packet (e.g., for control of NIC-level staged functionality).
- *What are the semantics of network traffic for required response by an edge device?* The research challenge is to collect traffic and be able to decode the protocol semantics for individual flows. By identifying packets that require no response, or only a trivial response (e.g., as could be handled by a proxy), the characteristics of idle periods may be changed to make power management at all times scales more fruitful. Being able to minimize needed wake-ups of a sleeping device is important to reducing energy use.
- *If applications can have knowledge of power state, can the amount of sleep time be extended?* We need to consider if new protocols, or extensions to existing protocols (e.g., to ICMP and SSDP in UPnP), for power

management notification would be useful to applications in determining the sleep state of peer devices. Having power management notification across a network (which is not at all possible today) may enable applications to defer or redirect requests and thus extend sleep time for edge devices. Cross-layer protocol adaptations to allow an application to identify and verify a networked device’s power management capabilities will enable application designers to explore new energy-saving techniques.

- *Can predictive algorithms be used to exploit the idle and low-demand periods for power management at multiple levels?* Predictive algorithms are needed to shutdown and wake-up an entire device (e.g., desktop computer) or network components within a device (e.g., link rate or function stages in a NIC). What estimators are the best for predicting the duration of future idle periods? Is there correlation between idle periods that can be exploited? What penalty functions should be used for under- and over-prediction of idle periods? There are many research challenges in predictive schemes for power management at multiple time levels including reducing energy use of high-speed links and improving the energy efficiency of NICs.
- *Within a NIC, how can dynamic assignment of hardware resources (e.g., via staging or reconfiguration) be used so that minimum energy is consumed for handling of various transactions as they arise?* Can enabling and assignment of these resources in a spatial or temporal fashion be used to adaptively react to network activity and conserve energy? What are the optimum tradeoffs in performance vs. efficiency? Can lower-power finite state machines (FSM) be used in place of dedicated processor logic? Can reconfigurable FPGA technology be fruitfully used for energy-efficient NIC design? What options are there for caching and/or using slower low-power memory on a NIC to improve energy efficiency?

2.2 Scope and significance of this research

The *scope* of this three-year project is to investigate methods to significantly reduce power consumption of Internet-connected desktop computers and other edge devices. The extent of the research will be to integrate power management across the hardware, instruction, protocol, and system layers focusing on desktop computers and embedded systems at the edge of the Internet. The *significance* of this project is enormous in the area of impact to society. **Simply “fixing” the wake-up and message response problems will result in higher adoption of power management in PCs that are currently shipping.** Having solutions to open power management problems will result in measurable energy savings for the individual user and society as a whole. Our research will enable the EPA to write a stricter specification for Energy Star for desktop computers [17]. This project will also increase the energy efficiency of future embedded systems and identify new directions and priorities in networking research.

3 Background and Related Work

This section describes existing power management and wake-up capabilities, and related work in improving the energy efficiency of the Internet.

3.1 Existing power management and wake-up capabilities in desktop PCs

Existing commodity PCs and the Microsoft Windows operating system support power management. The Advanced Configuration and Power Interface (ACPI) [2] is an industry standard for PC operating systems that addresses power-management interfaces. ACPI prescribes multiple levels of power-down, from a fully-powered system to intermediate states with one or several components powered off (e.g., a disk drive can be turned off after a period of inactivity). Very low power sleep states are defined in ACPI. PCs in a sleep state can use Wake on LAN (WOL) [56] Ethernet NICs to trigger a wake-up of the PC. WOL 10/100-Mb/s Ethernet NICs are readily available at prices of about \$6 [19]. Figure 3

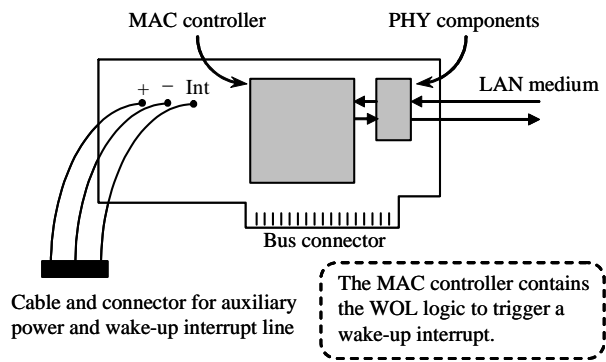


Figure 3. WOL Ethernet NIC with auxiliary power

shows a WOL NIC that has three functions:

1. Connection to auxiliary power to allow NIC operation to continue when the PC processor, memory, storage, and bus power-off (with the release of PCI 2.2, power management signals can be carried on the PCI bus).
2. Connection to a wake-up interrupt line, which is external to the PC system bus.
3. The ability to receive all packets when operating on auxiliary power and recognize a special WOL packet. On recognition of a WOL packet, a wake-up interrupt signal is generated.

A WOL packet is defined as any packet that in the data field contains the NIC MAC address repeated sixteen times. Since the WOL packet requires use of a MAC address, it is difficult to use it across an IP-routed network. This special packet is not part of the existing TCP/IP connection packet flows or semantics. Due to WOL packets being non-routable and not a standard, WOL has not achieved significant use. Most Dell PCs now ship with power management capabilities, auxiliary power, and WOL-enabled Ethernet NICs [39]. Motherboards from Intel ship with similar capabilities [30]. Some Intel NICs also support directed wake-up where packets containing the local IP address and/or a specific pattern can trigger a wake-up. Directed wake-up causes unnecessary wake-ups in some cases and fails to trigger needed wake-ups in other cases. **What is missing is the support to keep existing power management capabilities enabled and/or prevent unnecessary wake-ups – we address this open problem.**

3.1.1 Application and user disabling of power management

It is very important to understand why power management is disabled in existing desktop computers. Most PCs ship with power management enabled, thus the disabling must be intentional. In the past, power management was often disabled because it was a significant annoyance to users. With faster power-up times this annoyance is disappearing, however new reasons to disable power management are emerging with the greater dependence on “the network” for productive work. We have identified cases of both power management “breaking” applications and the opposite where applications “break” power management. The key issues are *device reachability*, *maintenance of permanent connections*, *knowledge of power state*, and *lack of prediction* to be able to power down without affecting user performance. Examples include lack of response to ARPs causing unreachability, applications with permanent TCP connections requiring periodic response, distributed discovery protocols failing due to lack of response, and use of fixed inactivity time-out to enter a low-power sleep state not adapting to actual application and user behavior.

3.2 Related work in improving the energy efficiency of the Internet

A very significant work on energy consumption of the Internet was Gupta and Singh’s ACM SIGCOM 2003 paper [25]. In [25] it was calculated that in 2000 networking devices consumed about 6 TWh of electricity and that this value was expected to increase by another 1 TWh by 2005. It was proposed that network interfaces in routers and switches can be put to sleep during idle times to reduce power consumption. Changes in routing protocols would need to be considered to achieve this. In [26] Gupta and Singh proposed power management capabilities for LAN switches. It was shown that a LAN switch that can enter a sleep state and be woken from it by packets queueing in a buffer (the buffer memory is not powered-off) can result in significant energy savings. The ideas presented in [26] are now funded by the NSF [50].

Server clusters in data centers have square-foot power demands that are two magnitudes greater than those of commercial offices and thus can be a localized power “hot spot” [45]. Significant work has been done in understanding how to trade-off response time and energy use in server clusters by enabling and disabling mirrored servers in a cluster [6, 10, 11]. When the request arrival rate is low, some servers in a cluster are powered-down. When the request arrival rate increases, powered-down servers are powered-up and brought back into full operation using the WOL mechanism. The NSF is funding work in this area [7].

The area of dynamic power management (DPM), a key concept of the “smart” NIC that we propose herein, is becoming increasingly important in the field. As described by Benini et al. in their survey paper on system-level DPM [4], the goal is to achieve performance at or near peak levels while having a proportionally low energy intake. DPM works by dynamically turning off or lowering performance of components not in full use at the time. For example, Gurumurthi et al. [27] introduced the concept of dynamic RPM for disk drives, whereby a controller, using a reactive policy, adapts hard disk speed to achieve energy savings with negligible performance loss. The basic concept is that by servicing requests at a lower RPM when performance is not critical, and increasing RPM when high performance is needed, the result is significant energy savings with little or no impact on the users.

One of the PIs (Christensen) has explored the energy efficiency of the Internet from a perspective of edge devices. In [32] the idea of a “Green TCP/IP” was first proposed in the context of a new energy-aware connection

semantic. In later work, such as [13] and [14], the need for further investigation and the potential for large energy savings are described. The potential of proxying to reduce the need for frequent wake-ups of desktop computers was first described in [12] and further studied in [13]. Several start-up companies are beginning to address power management of IT operations [1, 55]. Verdiem [55] markets a centralized energy management software product that controls the power management settings in PCs within a company. This centralized control can only address some of the problems causing power management to be disabled, can only work in large commercial operations, and cannot address energy efficiency of active devices. The existence of these companies and products is evidence of the possible economic benefits to individual users of increased use of existing power management.

4 Research Plan

This section elaborates on the research questions from Section 2 and describes how we will address them.

4.1 Research challenge #1 – characterization of idle periods at multiple time scales

The common thread to all of the research challenges is an understanding of the characteristics of idle periods in packet traffic as seen by an edge device. We seek a good understanding of the characteristics of idle periods in order to be able to exploit these periods for low-power sleep periods. Previous work in studying the characteristics of idle periods include [34, 40, 58], and more recently [3]. The results in the classic Jain and Routhier “packet trains” paper [34] show that inter-packet times are in general not exponentially distributed, Paxson and Floyd [40] show that human-generated interarrival times (e.g., interarrival time between telnet requests) are exponentially distributed, and Willinger et al. [58] and Barford and Crovella [3] show that packet traffic is self similar and thus packet interarrival times are often heavy tailed. This previous work has been focused on the eventual queueing behavior (e.g., for buffer sizing and QoS prediction) of packet traffic. **Our focus is different – we seek to understand idle periods from a perspective of how they can be exploited for power management – and thus additional work is needed in this area.**

For the USF dormitory traffic described in Section 2, we plotted the cumulated idle time between flows of these inter-flow idle times (figure 4). It can be seen that a large amount of idle time is in long idle periods. For the long idle periods (of many seconds or minutes), the entire PC can be put in a low-power state. We also found there to be a wide range of autocorrelation between idle periods. In some traces there was considerable autocorrelation in the idle periods for many lags, for other traces there appeared to be little or no autocorrelation. Correlation needs to be carefully studied since it will have implications on finding effective policies for predicting idle periods.

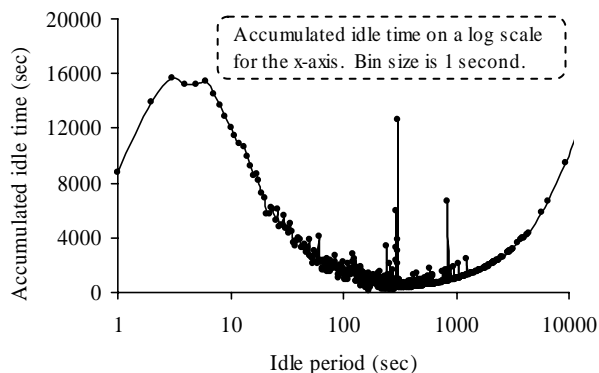


Figure 4. Accumulated idle time for USF PCs

We need to extend this work for many more traffic traces to include both commercial and residential networks running modern file sharing, chat, and video-related applications (whose traffic characteristics have not yet been well studied). We need to focus on the application mix present in the traffic traces. We also need to consider idle periods at inter-packet time scales, and not only inter-flow time scales as shown in the above figures.

In summary, for research challenge #1 we will:

- Collect traffic traces at the packet and flow levels from commercial and residential networks.
- Characterize idle periods at all time scales (such as inter-packet and inter-flow).
- Gain a better understanding of how prediction of future idle periods could be achieved.

4.2 Research challenge #2 – minimizing needed wake-ups by proxying and smart wake-up

When a desktop computer or other edge device is powered-down it “disappears” from the network. Failure to respond to DHCP, ARP, and other standard protocol messages will result in significant problems (such as losing its IP address) when the computer awakens at some later time. In addition, any current TCP connections will be

dropped by the server. We believe that many application and protocol-related messages that require a response do not require the full resources of a PC or edge device to generate the response. For example, ARP is a very lightweight protocol that could be run on a very small processor. A shortcoming with existing wake-up based schemes is requiring a desktop PC or other edge device to fully wake-up for useless or minor tasks. It should be possible to be more selective in wake-up and to proxy for some tasks normally performed by the PC operating system (which requires the PC to be fully powered-on). We will explore the possibilities of improved wake-up and proxying for responding to “network chatter.”

4.2.1 *Proxying to reduce the need for wake-up and provide new power management functionality*

An idle PC receives packets on its link at all times. The packets can be categorized as:

- *No response required* – packets that require no actions or response and are discarded by the protocol stack in the operating system. This includes broadcast bridging and routing protocol packets. This also includes “hacker” traffic such as port scans.
- *Minimal response required* – packets that require minimal action or a response that a proxy could handle. This includes ARP, ping, DHCP, and discovery packets.
- *Wake-up required* – packets that require operating system or application-level response. This includes TCP SYN packets for connection requests to applications with listens and SNMP GET requests.

In addition, some network protocols generate packets from a client. For example, DHCP lease renewal requests are generated locally from a PC or other device holding a DHCP-granted IP address. Packets from the first two categories (no response required and minimal response required) we call “network chatter.” As a preliminary experiment, we collected a packet trace from an idle PC networked by switched Ethernet in the engineering building at USF. In 12 hours and 40 minutes, 296,387 packets were received (this is over 6 packets per second). The NIC was configured to wake-up the PC upon receiving either a WOL packet or a directed packet. The system was configured for 10 minutes of inactivity time before transitioning into the Windows XP “standby” (low-power sleep) state. ARP packets constitute the slight majority of received packets and along with Universal Plug and Play, routing, and bridge protocol packets comprise about 80% of the total. The PC was fully powered-on for over 30% of the time due to ARP wake-ups.

Protocol proxying implemented on a NIC or within a first-level LAN switch could handle most network chatter eliminating the need to wake-up the full system for trivial (network-related) tasks. A NIC could include a small processor and software that would, when the PC is in a low-power state, act as proxy for the PC protocol stack and applications. In 2005 the additional cost to an Ethernet NIC of adding proxying capability via a small microcontroller is only a few dollars. This is about one order of magnitude less than the yearly electricity cost for a PC that is fully powered-on all the time. The proxy would filter packets that require no response, reply to packets that require a minimal response, and only wake-up the system for packets requiring a non-trivial response. The proxy would also generate packets for simple protocols such as DHCP based on a timer interrupt. Interfaces need to be defined to pass state information from a PC to its proxy in a NIC or LAN switch. This state information would include the IP address and which TCP ports are open for listening. With proxying, 91% of the nearly 300,000 packets traced in our preliminary experiment could have been filtered or trivially handled by a proxy. The remaining 9% of packets included TCP SYN packets, most intended for non-existent ports. The incoming ARP requests and ICMP packets can be handled by a proxy. The SYN packets require a response from the system only when there is an executing application with an open port matching the request.

Maintaining existing TCP connections is another challenge for the proxy. An idle TCP connection typically sends keep-alive messages for at least two hours with connection sequence numbers that have already been acknowledged by the receiver. Upon receiving these already acknowledged sequence numbers, the receiver – or its proxy – should transmit a duplicate acknowledgement for the last data received from the sender. This requires additional state knowledge to be transferred to a proxy. Messages that arrive within a TCP connection also need to be considered. These application messages may require a wake-up of the PC, or it may be possible to queue them (e.g., within the proxy) for later delivery to the PC. We plan to implement a proxy initially in emulated form on a standalone PC (similar to what was demonstrated in [13]) and finally in a prototype Ethernet NIC (see Section 4.5).

A proxying capability on a NIC can also yield entirely new functionality for power management. A proxy could participate in new protocol semantics intended for distributed power management as described in Section 4.3.

4.2.2 Smart wake-up to reduce unnecessary wake-ups and power-up time

A sleeping device should only awaken for a valid request that requires the resources of the full system. With proxying, described in the previous subsection, many requests may not need the resources of the full system. Wake-ups for invalid requests result in excessive time in a fully powered-on state. A valid request can be an incoming packet with a destination to an application in the device (e.g., a TCP SYN packet to an open port) or an internal timer trigger request (e.g., a timer for DHCP lease renewal request). The key challenge is how to determine if an incoming packet is destined for a valid application. Thus, the NIC must be “smart.” Transfer of state information between the main system and the wake-up function implemented in a NIC must be possible. What should this state information be? How should a wake-up function be implemented to use this state information? Should security be considered by only allowing authorized addresses to wake-up a device (i.e., an access list is maintained as part of the wake-up capability)? Existing wake-up functions in Ethernet NICs are based on WOL, wake-up on directed ARP, and limited pattern matching on received packets. We will explore what additional functions are needed to achieve a wake-up that only and always triggers on valid requests.

Finally, we will characterize the reduction in fully powered-on time that can be achieved from smart wake-up. Initially, this can be done by using packet traces. By the end of our project, we will evaluate smart wake-up in a real office environment using the prototype NIC described in Section 4.5.

In summary, for research challenge #2 we will:

- Investigate and construct a formal taxonomy of protocol messages and their effects on system state.
- Explore passive and active methods so that a device may determine when proxying is needed.
- Prototype and evaluate proxying capability to reduce the need for wake-up of an edge device.
- Explore new wake-up functionality to improve upon existing WOL and pattern match wake-up.
- Characterize the reduced power-on time, and thus energy savings, achievable from smart wake-up.

4.3 Research challenge #3 – power management notification by new protocol semantics

Power management semantics need to extend between devices and not just within a device, as is the case today. **By extending knowledge of power management between devices, overall sleep time may be increased and effects to user response time also minimized.** For example, a device requesting service (e.g., to download a file from a peer desktop computer in a P2P overlay network) could use knowledge of the power management state of the desktop computer to make decisions such as whether:

- 1) Sufficient resources are powered-up to service the request.
- 2) A wake-up is scheduled in the near future and the request could wait for this wake-up to occur.
- 3) There is another device in an appropriate power state to which the current request could be redirected.

Another example of energy savings possible by new protocol semantics would be where a media device is playing a media stream sent by a home media center PC. Rather than continuously stream the data from the PC, the device could wake-up the PC periodically and “grab” sufficient media data for several minutes and let the PC transit to a low-power sleep state. When establishing the connection, the PC would inform the device of its power management capabilities and enable this batched mode of operation.

Power management notification and control across a network has the potential to create security vulnerabilities at many new levels. For example, if an outsider can detect that a home PC has been in a low-power sleep mode for a long period of time then a burglar could use this information to determine if a house is unoccupied [41]. Also, if the power state of a device can be changed across a network (e.g., via a wake-up message), then the potential exists to “crash the power grid” by simultaneously waking-up all the power-managed network devices in a large building [41]. We need to carefully consider both well known and entirely new security issues in our investigation of proxying for existing protocols and design of new protocols for power management notification.

For this research challenge we will investigate whether ICMP can be extended to include power management notification and control for all devices that support IP. We will also look at Universal Plug and Play (UPnP) [57] to understand how power management notification semantics could enable the use of power management in UPnP devices. Currently, UPnP is “power-unaware” and its distributed SSDP discovery protocol requires all devices to be fully powered-on at all times. Fully distributed protocols and applications almost always rely on all devices being fully powered-on at all times (e.g., P2P overlay networks with their permanent TCP connections). We will investigate power notification semantics at the application level.

In summary, for research challenge #3 we will:

- Investigate new protocol semantics to extend power management notification from intra-device (e.g., as implemented by ACPI) to inter-device.
- Explore the implementation of such new semantics in ICMP, SSDP in UPnP, and other existing protocols.
- Explore the need for power management semantics in network applications.

4.4 Research challenge #4 – adaptive link rate for Ethernet and predictive policies

New mechanisms are needed to scale-down energy use during low utilization periods. Predicting and exploiting idle and low utilization periods is key to effective power management. Predictive policies need to be investigated.

4.4.1 Adaptive link rate for Ethernet

As Ethernet has increased its data rate to 1 Gb/s and emerging 10 Gb/s the energy consumed by a NIC has increased. Existing 10/100 Mb/s Ethernet NICs consume about 1 W [44], 1 Gb/s Ethernet NICs consume about 7 W [37], and emerging 10 Gb/s NICs can be expected to consume far more (we measured 15 W for one brand). On existing 1 Gb/s NICs the power consumption is dropped by at least 2 W (by our measurements on three brands of NICs) when the data rate is reduced to 100 Mb/s. A reduction of at least 4 W is thus possible if the link between a desktop PC and its LAN switch is operated at 100 Mb/s instead of 1 Gb/s (i.e., savings occur in both the desktop PC and switch NICs). The energy consumed by a NIC at data rates of 100 Mb/s and higher is approximately constant whether the link is transmitting and receiving packets or is idle due to PHY layer signaling. When transitioning to Windows XP standby mode, some NICs already drop the data rate from 1 Gb/s to 10 Mb/s to save energy [31]. From this we are motivated to think about the possibility of adaptive link rate and ask two questions:

1. What percentage of time would a lower Ethernet link data rate yield the same QoS, or user performance (e.g., measured in user-perceived delay), as would a higher data rate?
2. Is it possible to transition between Ethernet link data rates while the link is active and reduce energy consumption without a user-perceived delay or other impact to network QoS?

The link for desktop to LAN switch for a single user is likely to operate at low utilizations for much of the time [38]. The high data rate of an Ethernet link is primarily intended for burst applications, such as infrequent very large file transfers. Thus, there may be a considerable opportunity for energy savings in adopting an adaptive link rate. We will investigate an adaptive link rate mechanism that can be used to match the data rate of an Ethernet link to the link utilization. With an appropriate policy to use this mechanism, we believe that link data rate can match the traffic demand without any user perceivable added delay and achieve a large savings in overall energy consumption. Simulations using simple utilization-based reactive policies on traces from the USF campus network show significant savings potential. Start-up time for a link is conservatively 10,000 clock cycles [51]. We thus estimate that it would be possible for copper Ethernet link data rates to be dynamically changed in about 1 millisecond. Existing Ethernet NICs capable of operating at multiple data rates use a mechanism called Auto-Negotiation to determine at which data rate to operate. For 1 Gb/s Ethernet devices, completing this process takes a minimum of 256 milliseconds due to backward compatibility issues with 10/100 Mb/s Auto-Negotiation. Thus, using the existing Auto-Negotiation capability to change the data rate dynamically would disable the link for 100's of milliseconds. This would result in unacceptable impacts to QoS. **We will explore new MAC and physical layer approaches for fast negotiation and triggering of adaptive link rate changes.**

4.4.2 Predictive policies for system power-down and adaptive link rate

We will investigate both reactive and proactive (predictive) policies for system power down (at the level of the full PC or device and also for components within the device – see Section 4.5) and for the new adaptive link rate mechanism. A reactive policy may be feasible if the impact (in added delay and/or packet loss) to a user of a power-up or data rate transition is minor. However, in many cases a predictive proactive policy will be needed to minimize transition delay effects (e.g., of power-up) to the user. Significant work has been done in studying prediction for network round trip time [33], disk drive idle time [20, 21], service time for load balancing [28], and event-driven applications [29, 46, 52]. Methods studied for dynamic power management [36] include adaptive methods [29, 52],

Markov decision process based strategies [5, 15, 42, 49], adaptive learning trees [16], and generalized stochastic Petri nets [43]. In [29] Hwang and Wu study exponential smoothing as a means of predicting system idle periods and an improved “watchdog” exponential smoothing method that better handles prediction of occasional long idle periods is proposed and studied. We have experimented with a modified exponential smoothing method (see figure 5) suitable for prediction of when to shutdown for idle periods.

1. if (beginning of idle period)
2. power-down
3. $T1 =$ predicted duration of this idle using exponential smoothing
4. wait for $T1$ seconds to expire or detection of an arrival event
5. power-up
6. if (no arrival event has occurred in past $T1$ seconds)
7. $T2 =$ estimated 95% median value of idle periods
8. wait for $(T2 - T1)$ seconds to expire or detection of an arrival event
9. if (no arrival event has occurred in past $T2$ seconds)
10. power-down
11. wait for detection of an arrival event
12. power-up
13. else
14. power-up

Figure 5. Threshold-based predictive power-up policy

At the start of an idle period the system is powered-down and the duration of the idle period is predicted using exponential smoothing (line 3 in figure 5). At the end of the predicted idle duration the system is powered-up (if not already powered-up by an arrival event) and a threshold decision, based on an online calculation of the 95% median idle time value, is made to predict if the current idle period will be long in duration. Our method exploits long idle periods. Figure 6 shows the improvement in percentage of idle time used for power-down with our new method compared to only using exponential smoothing for prediction. The bars in Figure 6 show the percentage of idle time exploited for power management (i.e., used for power-down sleep time) and the lines show the percentage of arrivals that find a powered-down system and suffer a power-up delay penalty. Figure 6 was generated using a 30 minute packet trace of the 10th busiest PC on September 21, 2004 on the USF campus (the trace comprises over 900,000 packets totaling 570 MBytes sent and received). It can be seen that our method gives an approximate 1.5x improvement in idle time used for power-down at only a very small expense in power-up delay penalty. Some traces give even better results. The results shown here are for idle periods and complete system power-down. We will investigate the suitability of similar methods for the different case of predicting and exploiting low utilization periods using adaptive link rate. For idle and low utilization periods with non-zero autocorrelation (a case we believe is fairly typical since we have measured non-negligible autocorrelation in many packet traces) we believe that even more effective predictive methods can be found. **An autocorrelation signature could be exploited in a predictive policy.** The impact of wake-up on power use and on the response time to service a request needs to be evaluated. This impact can determine what levels of power management can be reasonably achieved for a given QoS trade-off. Predictive policies will be investigated for both complete power-down to a low-power sleep state of the edge device (question #1 in Section 1.1) and adaptive link rate to conserve energy while in active use (question #2).

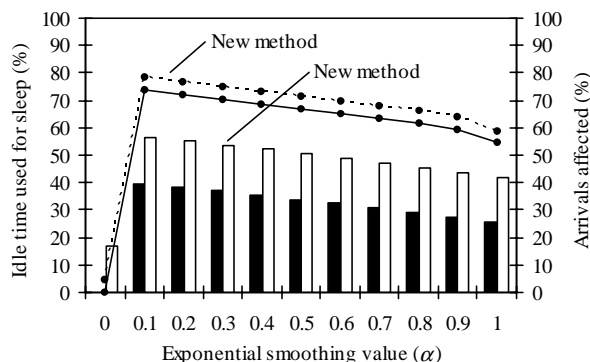


Figure 6. Sleep time and arrivals affected

In summary, for research challenge #4 we will:

- Architect a new fast Auto-Negotiation mechanism for Ethernet to adaptively change link data rates to improve energy efficiency with minimal impact on QoS.
- Investigate both reactive and proactive (predictive) policies for both complete system power-down and for adaptive link data rate. Study prediction methods for autocorrelated idle periods.

4.5 Research challenge #5 – architectures for energy-efficient, multistage, adaptive NICs

As discussed and implied in the preceding sections, the role of the NIC within the edge device (e.g., desktop PC), or alternatively in its first-level LAN switch, will be critical in achieving energy efficiency of the system. Although traffic is almost constantly being received by any idle edge node on the network, much of this traffic is network chatter that we propose could be filtered when possible or handled by a “smart” NIC, both without awakening the

host. Even more challenging, the potential exists for the proxy to handle responses to critical traffic that might otherwise require the host, such as maintaining an open but idle TCP connection. Concomitant with research goals for protocol proxying, smart wake-up, distributed power management between nodes across the network, and adaptive use of data rate on network links, all based on traffic characterization and predictions, is the requirement for a versatile, adaptive, and energy-efficient NIC. The research challenge here is to investigate, quantify, compare, and then determine the optimum architecture to address these challenges. **We propose to focus upon fixed or reconfigurable levels of staged hardware functionality to realize energy-efficient operation via adaptive spatial or temporal assignment of hardware resources to network transactions with power, performance, and functionality scaling with hardware-level dynamic power management.** Tradeoff analyses will be done to determine the relationship between increased functionality in the smart NIC versus overall energy savings and cost.

The potential for power savings with the NIC itself is interesting and important, but far more benefit can be gained by powering down the surrounding system via a “smart” NIC as much and as often as possible, which is the primary focus of this project. As described in the previous section, energy use has increase from about 1W for a 10/100 Mb/s NIC to about 15 W for a 10 Gb/s NIC. There is clearly room for improved energy efficiency within the NIC itself. Extending and applying ideas from ACPI, we divide the NIC into two or more fundamental levels of functionality. Each level will have a distinct tradeoff between power, performance, size, and functionality. The levels and their power/functionality tradeoffs will be analogous to ACPI such that at each device state, different levels or combinations of components on the interface will be functional. This is functionality specialization for power consumption.

Figure 7 shows our proposed concept of a multistage controller. The lowest level of functionality corresponds to a simple Finite State Machine (FSM) device with very limited functionality, power consumption, cost, and area. The function at this “sentry level” is simply to maintain the network connection while allowing higher levels to remain in low-power standby mode. This level can be implemented using several different technologies such as ASIC or FPGA. Although historically FPGAs have not generally been considered first for low-power solutions, new devices (e.g., Xilinx Spartan-3L) show promisingly low levels of per-slice energy consumption, and the hardware-reconfigurable nature of FPGAs provides a promising approach to achieving an ideal functionality-to-energy ratio and warrants further investigation. Here, a set of configuration files could be co-located and loaded into the FPGA on the fly as demands dictate, subject to overhead in reconfiguration time. Thus, at a given level in this hierarchy, the subsystem could temporally adapt the use of hardware resources (and thereby energy resources) to meet the dynamic demands of traffic. The clock rate of the reconfigured hardware resources of the FPGA can also be dynamically adapted to trade performance for energy savings. By contrast, the efficiency, compactness and economy of scale of an ASIC is much more clear, and commercial devices (e.g., TCP/IP/Ethernet controllers) exist on the market. We will explore the power consumption tradeoffs between reconfigurable FPGA and purely ASIC designs. Using results from previous phases, we will identify the optimal tradeoff between functionality and power for each level. For example, a “sentry level” device may be able to handle simple requests that are commonplace in network housekeeping such as DHCP, or it may only be a logical circuit to wake the next level up to handle any requests addressed to the host being served.

In this multistage framework, each higher level would expand upon the functionality, capability, and performance of the lowest level. Its purpose would be to service the majority of routine network transactions not satisfied by the lower level(s) while still achieving power savings by handling common requests without awakening upper levels. Hardware resources could be dedicated for each level of functionality, as illustrated in figure 7, and each level adaptively enabled or disabled (i.e., powered on or off) as traffic patterns and thereby the control algorithms dictate. Alternately, a single FPGA could serve this same purpose via run-time reconfiguration. In either case, the hardware must be smart enough to sleep when not being used and respond in a timely manner. Concomitant with the increased power allowance at higher levels, a low-power network processor (NP) could also provide a promising approach at higher levels, depending upon the size of the system, with flexibility in power consumption in terms of dynamic voltage level and clock frequency scaling for tradeoffs in performance and energy. As shown in figure 7, several such levels above the lowest level could exist in terms of a function cache where

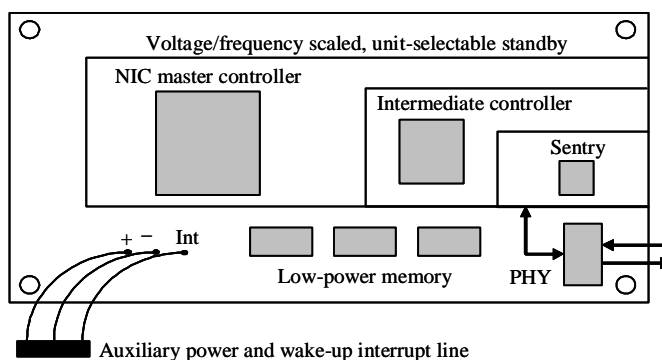


Figure 7. Ethernet NIC with a multistage controller

higher levels support more network transactions but at a higher cost in energy efficiency, up to at most a complete TCP/IP/Ethernet controller that offloads the host (desktop) processor or the central (embedded) processor entirely.

The software developed with this hardware will involve and support the interactions between the NIC and the host and its operating system. The software must provide a communication interface so that the operating system can influence the power mode of the device according to dynamic and/or user configurations. It must also allow the network interface to inform the operating system of a power state switch since the controller will be self-managing – a very important feature in edge devices, such as embedded systems that do not have robust operating systems. These communications will conform to the ACPI interface for those systems that adhere to this standard.

Another facet in the architecture relates to memory options in order to determine the optimum configuration. The type, speed, and capacity of memory allotted (if any) to each level will be analyzed in terms of power, performance, and area constraints. From the knowledge gained from research challenges #1 and #2, we will classify the most common requests and consider a simple transaction caching scheme for servicing these requests. A cache can store the responses from the higher levels for simple repetitive requests that do not require modification or access to host resources such as main memory or disk. This cache might enable the lower levels to preserve the network connection without higher levels. Other uses of low-power memory at one or more levels could include storage for configuration files for reconfigurable FPGAs, power management control for multiple levels, DMA transfers, main memory mapping, or even reflective memory for transaction state. Several of these options would allow the NIC to bypass the central processor and access main memory to handle a TCP request. The aforementioned issues will be explored and tradeoffs evaluated with the prototyping and analysis of an energy-efficient NIC as an extension to the work at the end of research challenge #4. A PCI-based prototype will be constructed and used to conduct experiments and gather data to determine where and how multiple independent stages of functionality and control, operation caching, etc. can best support network transactions and thus permit the host (or higher-level unit on the multistage NIC) to sleep more often and for longer periods so as to minimize the overall energy consumed. While focused on the desktop PC as the dominant edge device of today's Internet, concepts developed and insight gained in this research will be extendable to meet the needs of embedded systems where power consumption for communication may outweigh all other factors.

In summary, for research challenge #5 we will:

- Investigate energy-efficient, adaptive architectures for Internet edge devices featuring the layering of functionality with staged standby modes – the “smart” NIC.
- Explore energy efficiency, functionality, and performance tradeoffs between reconfigurable FPGA (temporal) and ASIC (spatial) multistage strategies for proxying and smart wake-up.
- Explore memory architectures and options for low-power memory to increase transaction hit rate of an energy-efficient NIC.
- Prototype and evaluate an energy-efficient NIC with multistage functionality and energy profiles.

4.6 Research application – prototyping and evaluation

The culmination of this research project will be a comprehensive set of experiments and detailed analyses conducted on a full prototype of the “smart” NIC that we will design and implement as a PCI card. This NIC will adaptively support all of the functionality and energy-conserving options explored in the five areas of research challenge described above and be adaptively driven by the reactive and predictive control algorithms developed in the research. A comprehensive evaluation plan will be undertaken including a broad range of energy and performance measurements on both the NIC and its offloaded host, first in an environment where the prototype “smart” NIC is stimulated by a broad variety of traffic patterns driven by network instrumentation and various Ethernet switches and servers in a controlled lab environment, and then in a production environment with the NIC and host operating as a typical edge node on the Internet.

Prior to the development and comprehensive testing of the full prototype, a series of initial prototype designs, each supporting a subset of the functionality of the “smart” NIC, will first be investigated. First, we propose to use CAD simulation tools to build virtual prototypes for each of the functionality options (e.g., sentry unit to filter or respond to simple network chatter) and study the energy and performance characteristics and tradeoffs of various approaches (e.g., FPGA, ASIC, and NP). Next, we will build several standalone small-scale prototypes each with limited functionality to further evaluate the most promising architectural options for the “smart” NIC. These early prototypes will help ensure that the final, full prototype meets all of the requirements to support the mechanisms investigated in this project.

4.7 Management plan and project milestones

The connection between the University of South Florida (Christensen) and the University of Florida (George) is synergistic. Christensen has expertise in traffic characterization and has completed initial work on the idea of proxying [12, 13]. George has expertise in network architectures, hardware-reconfigurable computing, and hardware design, including energy-efficient designs [24]. Christensen and George have worked together on several conference committees and have organized workshops – a good working relationship has already been well established. This project will require four PhD students, two at each university, for three years. The two students at USF will be responsible for challenges #1, #2, #3, and part of #4. The students at UF will be responsible for challenges #4 and #5, and be responsible for the design and completion of a prototype NIC that will incorporate the results from all of the challenges in this project. The design and completion of the prototype NIC falls within the expertise of George at UF. Clearly, all of the research challenges are interrelated and significant cooperation and communication between the PIs and their students will occur. Email, telephone, and video conferencing facilities will be used on a weekly basis to maintain communications between USF and UF. We will also plan for periodic face-to-face meetings to be held alternately in Tampa (USF) or Gainesville (UF). The expected milestones are:

- *End of year #1* – We expect to have traffic characterization for research challenge #1 completed and have a fully working software prototype of proxying for research challenge #2 completed. Initial design of a prototype NIC will be completed for challenge #5.
- *End of year #2* – We will continue to refine our ideas on proxying and smart wake-up. We will have made progress on defining new protocol semantics for research challenge #3. We will also have initial results in research challenge #4 – at least sufficient to begin dissemination to the IEEE 802.3. We will also have applied our findings from year #1 to predict overall energy savings.
- *End of year #3* – The prototype NIC embodying ideas from all the research challenges will be complete. Dissemination of results to the EPA, IEEE 802.3, and industry will be complete. A workshop will be organized and held, and course materials will be completed.

5 Expected Outcomes and Dissemination Plan

5.1 Expected research outcomes

Our expected research outcomes are a better understanding of idle periods in network traffic at multiple time scales and how power management can be achieved in network edge devices at all time scales of idle periods. Our four most significant expected outcomes are:

1. A solution to the existing “network problem” that causes power management features to be disabled in desktop computers and other edge devices. This solution is largely one that can be solved with proxying and smart wake-up capabilities on an Ethernet NIC or LAN switch. This solution will enable power savings of many TWh/yr in the next three to five years with significant economic savings and environmental benefit.
2. The first steps toward future protocols to support secure power management notification across a network.
3. A multiple time scale and level approach to power management applied to NICs for both desktop computers and future embedded systems. This will close the gap in power management where mobile systems and servers have already been addressed, but desktop computer and other edge devices have not been addressed.
4. Significant influence on industry directions and the EPA Energy Star specification for desktop computers and future edge systems (such as set-top boxes and other networked devices that may fall under Energy Star).

To support these outcomes, we expect to complete (and make generally available) a prototype energy-efficient NIC that supports proxying, smart wake-up, and adaptive link rate and functionality. We will also publish and present our findings at key networking conferences and journals, and we will organize and host a workshop.

5.2 Dissemination of research results to the IEEE 802.3 and industry

Key to the success of this project is disseminating the results. The prototype energy-efficient NIC that we will build will be useful in disseminating the new ideas from this research. In order to incorporate adaptive link rate in a future generation of Ethernet NICs in edge devices and first-level LAN switches, the IEEE 802.3 needs to consider exploring new ideas for fast autonegotiation (of link speed). Our research will likely result in two directions for fast autonegotiation – one method based on using PHY layer symbols and another method based on MAC frame exchange. We believe that the IEEE 802.3 will be receptive to our ideas based on their interest in standardizing a broad range of solutions from mobile, to desktop, to metropolitan use of Ethernet at data rates of 1 Gb/s and higher. We will work with industry to promote our results in proxying, smart wake-up, and adaptive link rate and functionality to be incorporated in future NICs.

5.3 Influencing the EPA Energy Star specification

The overriding influence on energy efficiency of computer equipment is the EPA Energy Star specification. The specifications written by the EPA dictate the priorities that manufacturers (and thus also standards groups) must take on improving energy efficiency if they wish to sell their products to the US Government. The federal government is a large customer of computer equipment and thus manufacturers are very motivated to comply with the EPA Energy Star specification. The EPA Energy Star program has also greatly influenced consumer purchasing. One of the major successes we expect from this project is to influence the EPA Energy Star specification. The current very low rate of use of power management on commercial-sector desktop PCs is a serious problem for Energy Star. EPA needs principles or standards to refer to in future specifications [18] that are technically sound that will greatly raise the enabling rate of power management and so increase program effectiveness and credibility. EPA lacks the technical resources and funding resources to develop this information – this project will address this critical need. **We have already established a relationship with the EPA – the attached letter of support from Craig Hershberg describes how we will influence the EPA Energy Star specification.**

5.4 Intellectual merit and broader impact

5.4.1 *Intellectual merit and expected impact to future research directions*

Significant progress has been made in power management at lower levels for mobile systems. Server power management has also been addressed. However, power management of the largest energy users – devices at the edge of the Internet including desktop computers and future embedded systems of all kinds – has barely been addressed. There is considerable intellectual merit in addressing power management for these devices. This research will define the importance of energy consumption of network edge devices as an economic and environmental issue. Dynamic power management and network protocols need to be investigated synergistically and not as separate entities and this must be done within the scope of both existing and future Internet applications and protocols. We plan to organize and host a workshop on this theme to further seed research interest in this direction. We thus believe that this research will have a very significant intellectual impact to long-term future directions and priorities in networking research.

5.4.2 *Broader impact to society*

The broader impact of the research to society will be significant as measured in energy conservation in the many TWh/yr (and the resulting decrease in greenhouse gases). A savings of 1 TWh/yr equals \$80 million at 8 cents per KWh. Conservation of energy from Internet devices is of growing urgency as the realization that the microprocessor and the Internet are major consumers of energy. This awareness exists at the very highest level of government. Executive Order 13221 [23] addresses standby power consumption and states a requirement of 1 W standby for all government-purchased electrical devices which have a standby mode. As described earlier in this proposal, we will work closely with government agencies (EPA) and companies that have the most influence on power consumption of Internet-connected devices. With the assistance of LBNL, expected energy savings will be quantified. This project will have near-term (3 to 5 years) impact for desktop computers and addresses outcomes not currently under development. In the next 3 to 5 years many households and businesses will get broadband connections (e.g., via cable modem or fiber) to the Internet and be “always on.” Beyond 3 to 5 years are new

devices such as television set-top boxes. If millions of households acquire set-top boxes, energy efficiency of these devices will be critical [47]. The included letters of support from LBNL and EPA attest to the expected broader impact to society measured in energy savings of billions of dollars per year.

5.4.3 Broader impact to education and outreach

Research and education are intertwined. The broader impact to education and outreach will include the following:

- Completion of four PhD students trained in an area of national and global importance. We expect one of the PhD students to be from an under-represented population addressing the great need for increased diversity in the future pool of researchers.
- Orientation of several undergraduates students to research (through yearly NSF REU supplements)
- Involvement of two K-12 teachers in summer research located at USF and UF (through NSF RET supplements). The teachers will develop lesson plans to be used in their schools.
- New course materials to bring a focus on power management into the undergraduate classroom. These course materials will be made available via the Web and will include at least one laboratory exercise.
- Influence on networks and performance evaluation textbook authors to include power management topics.
- Workshop on power management of network edge devices to seed and establish future research directions.

We urge the reviewers to consider the expected measurable broad impact to society of this proposal. Even if only partially successful, this project will prove to have been a very good investment by the NSF. The attached letters of support clearly show the expected economic and environmental benefits from this project.

6 Results from Prior NSF Funding

Christensen had prior NSF support (“CAREER: Performance Evaluation of Gigabit Ethernet Network, A Systems and Experimental Approach.” ANI-9875177, 1999 to 2003). George had prior NSF support as co-PI (“CISE RR: Collaborative Research on Wide-Area Network Computing using Virtual Machines,” EIN-0224442, 2002-2005; “MRI: Acquisition of CASTOR: A High-Performance Communication and Storage Backbone for Data-Intensive Scientific and Engineering Computing,” CNS-0421200, 2004-2007). The results from this support are:

- ANI-9875177 resulted in progress in the areas of new switch architectures for variable length packets and URL switching. Combined Input and Crossbar Queued (CICQ) switches were investigated and shown to be feasible to implement and better in performance than purely Input Queued (IQ) switches. Two graduate students – Zornitza Genova Prodanoff (US citizen) and Kenji Yoshigoe (Japanese national) – completed their PhDs in 2003 and 2004, respectively. Zornitza is currently an Assistant Professor at the University of North Florida, and Kenji is an Assistant Professor at the University of Arkansas at Little Rock. Twenty publications, two web sites, and one generally available simulation tool resulted from this grant. Two REUs and one RET broadened the impact of this grant.
- EIN-0224442 is an on-going equipment grant that is establishing a test bed for research in information grids linking the University of Florida (UF), Purdue University, and other collaborating institutions. Defining features of the test bed include virtualization capabilities, wide-area distribution, scalable capacity, and heterogeneity. Components include a powerful mainframe computer, compute clusters, and high-capacity storage units. A number of interrelated research projects have been enabled by this test bed with a goal of developing VM-based middleware for grid computing for virtualized end resources, monitoring and prediction, interactive computing, virtual file systems, data management, cycle selling, and security.

CNS-0421200 is a new equipment grant that will establish a new campus research network at UF. This 10 Gb/s Ethernet network will support research with the new HPC Center and campus research grid at UF by linking together several distributed sites of the center to one another, and to other key HPC facilities on campus, as well as enabling high-speed connectivity to Florida Lambda Rail, the new state-wide, high-bandwidth research and education network for the State of Florida. CASTOR will synergize and strengthen collaboration between multidisciplinary teams across the institution and provide a unique resource for national and international projects.

References Cited

- [1] "1E: Software Products: Power Management: Overview," 2004. URL: <http://www.1e.com/SoftwareProducts/PowerManagement/Index.aspx>.
- [2] Advanced Configuration And Power Interface Specification, Revision 3.0, September 2, 2004.
- [3] P. Barford and M. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation," *Proceedings of the ACM SIGMETRICS*, pp. 151-160, July 1998.
- [4] L. Benini, A. Bogliolo, and G. De Micheli, "A Survey of Design Techniques for System Level Dynamic Power Management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 8, No. 3, pp. 299-316, June 2000.
- [5] L. Benini, A. Bogliolo, G. Paleologo, and G. De Micheli, "Policy Optimization for Dynamic Power Management," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 18, No. 6, pp. 813-833, June 1999.
- [6] R. Bianchini, "Research Directions in Power and Energy Conservation for Clusters," *Technical Report DCS-TR-466*, Department of Computer Science, Rutgers University, November 2001.
- [7] R. Bianchini (PI) and U. Kremer (Co-PI), "CISE Research Infrastructure: Infrastructure of Power-Aware Server Clusters," *NSF Award Abstract - #0224428*, August 16, 2002. URL: <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0224428>.
- [8] K. Cameron (PI), "CAREER: High-Performance, Power-Aware, Distributed Computing," *NSF Award Abstract - #0347683*, February 6, 2004. URL: <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0347683>.
- [9] W. Carnoy, "My Digital Home's Dirty Little Secrets," *CNET Reviews*, July 9, 2004.
- [10] J. Chase and R. Doyle, "Balance of Power: Energy Management for Server Clusters," *Proceedings of the Eighth Workshop on Hot Topics in Operating Systems*, May 2001.
- [11] J. Chase, D. Anderson, P. Thakur, and A. Vahdat, "Managing Energy and Server Resources in Hosting Centers," *Operating Systems Review*, Vol. 35, No. 5, pp. 103-116, 2001.
- [12] K. Christensen and F. Gullede, "Enabling Power Management for Network-Attached Computers," *International Journal of Network Management*, Vol. 8, No. 2, pp. 120-130, March-April 1998.
- [13] K. Christensen, P. Gunaratne, B. Nordman, and A. George, "The Next Frontier for Communications Networks: Power Management," *Computer Communications*, Vol. 27, No. 18, pp. 1758-1770, December 2004.
- [14] K. Christensen, B. Nordman, and R. Brown, "Power Management in Networked Devices," *IEEE Computer*, Vol. 37, No. 8, pp. 91-93, August 2004.
- [15] E.-Y. Chung, L. Benini, A. Bogliolo, Y.-H. Lu, and G. De Micheli, "Dynamic Power Management for Non-Stationary Service Requests," *IEEE Transactions on Computers*, Vol. 51, No. 11, pp. 1345-1361, November 2002.
- [16] E.-Y. Chung, L. Benini, and G. De Micheli, "Dynamic Power Management using Adaptive Learning Tree," *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 274-279, November 1999.

- [17] "Computers: Energy Star," 2004. URL: http://www.energystar.gov/index.cfm?c=computers.pr_computers.
- [18] "Computer Specification: Energy Star," 2004. URL: http://energystar.gov/index.cfm?c=revisions.computer_spec.
- [19] D-Link DFE-530TX+ 10/100 PCI NWAY NIC w/ Wake On LAN, D-Link Systems Inc., 2003. URL: <http://shop3.outpost.com/product/2700333>.
- [20] F. Douglis, P. Krishnan, and B. Bershad, "Adaptive Disk Spindown Policies for Mobile Computers", *Proceedings of the Second USENIX Symposium on Mobile and Location-Independent Computing*, pp. 121-137, April 1995.
- [21] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the Power Hungry Disk," *Proceedings of the 1994 Winter USENIX Conference*, pp. 292-306, January 1994.
- [22] M. Elnozahy, M. Kistler, and R. Rajamony, "Energy Conservation Policies for Web Servers," *Proceedings of USITS, 4th USENIX Symposium on Internet Technologies and Systems*, March 2003.
- [23] Executive Order 13221, "Energy-Efficient Standby Power Devices," signed by President George W. Bush, The White House, July 31, 2001.
- [24] A. George, "High-Performance Computing and Networking System for Advanced Antisubmarine Warfare," Annual Report, Project # N00014-99-1-0278, Ocean Atmosphere and Space (OAS) Science and Technology Department, Office of Naval Research, December 2003.
- [25] M. Gupta and S. Singh, "Greening of the Internet," *Proceedings of ACM SIGCOMM*, August 2003.
- [26] M. Gupta, S. Grover, and S. Singh, "A Feasibility Study for Power Management in LAN Switches," *Proceedings of the 12th IEEE International Conference on Network Protocols*, October 2004.
- [27] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "DRPM: Dynamic Speed Control for Power Management in Server Class Disks," *Proceedings of the International Symposium on Computer Architecture*, pp. 169-179, June 2003.
- [28] M. Harchol-Balter and A. Downey, "Exploiting Process Lifetime Distributions for Dynamic Load Balancing," *ACM Transactions on Computer Systems*, Vol 15, No. 3, pp. 253-285, August 1997.
- [29] C. Hwang and A. Wu, "A Predictive System Shutdown Method for Energy Savings of Event-Driven Computation," *ACM Transactions on Design Automation of Electronic Systems*, Vol. 5., No. 2, pp. 226-241, April 2000.
- [30] Intel Desktop Board D815EEA – Technical Product Specification, *Order Number A16964-001*, May 2000.
- [31] Intel 82541PI Gigabit Ethernet Controller, 2004. URL: <http://www.intel.com/design/network/products/lan/controllers/82541pi.htm>
- [32] L. Irish and K. Christensen, "A 'Green TCP/IP' to Reduce Electricity Consumed by Computers," *Proceedings of IEEE Southeastcon*, pp. 302-305, April 1998.
- [33] V. Jacobson, "Congestion avoidance and control," *ACM SIGCOMM Computer Communication Review*, Vol. 18, No. 4, pp. 314-329, August 1988.
- [34] R. Jain and S. Routhier, "Packet Trains-Measurements and a New Model for Computer Network Traffic," *IEEE Journal of Selected Areas in Communications*, Vol. SAC-4, No. 6, pp. 986-995, September 1986.

- [35] K. Kawamoto, J. Koomey, B. Nordman, R. Brown, M. Piette, M. Ting, and A. Meier, "Electricity Used by Office Equipment and Network Equipment in the U.S.: Detailed Report and Appendices," *Technical Report LBNL-45917*, Energy Analysis Department, Lawrence Berkeley National Laboratory, February 2001.
- [36] Y. Lu, E. Chung, T. Simunic, L. Benini, and G. Micheli, "Quantitative Comparison of Power Management Algorithms," *Proceedings Design, Automation and Test in Europe Conference and Exhibition 2000*, pp. 20-26, March 2000.
- [37] "NP1800/NP1880 PCI Ethernet Gigabit NICs," Netcomm Inc., 2003. URL: http://www.netcomm.com.au/one/support/specification/NP1800_NP1880Gigabit_NIC_Screen.pdf.
- [38] A. Odlyzko, "Data Networks are Lightly Utilized and Will Stay That Way", *Review of Network Economics*, Vol. 2, No. 3, pp. 210-237, September 2003.
- [39] Optiplex GX60, 2003. URL: http://www.dell.com/us/en/bsd/products/model_optix_3_optix_gx60.htm.
- [40] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", *IEEE/ACM Transactions on Networking*, Vol. 3 No. 3, pp. 226-244, June 1995.
- [41] Personal conversations with Bruce Nordman at Lawrence Berkeley National Laboratory, April 2004.
- [42] Q. Qiu and M. Pedram, "Dynamic Power Management Based on Continuous-time Markov Decision Processes", *Proceedings of the 36th Design Automation Conference*, pp. 555-561, June 1999.
- [43] Q. Qiu, Q. Wu, and M. Pedram, "Dynamic Power Management of Complex Systems using Generalized Stochastic Petri Nets," *Proceedings of the 37th Design Automation Conference*, pp. 352-356, June 2000.
- [44] "Quick Spec: Compaq NC3123 Fast Ethernet NIC PCI 10/100 Wake on LAN," May 8, 2001. URL: http://h18000.www1.hp.com/products/quickspecs/10305_na/10305_na.HTML.
- [45] S. Rahman, "Power for the Internet," *IEEE Computer Applications in Power*, Vol. 4, No. 4, pp. 8-10, October 2001.
- [46] D. Ramanathan, S. Irani, and R. Gupta, "An Analysis of System Level Power Management Algorithms and Their Effects on Latency," *IEEE Transactions on Computer Aided Design*, Vol. 21, No. 3, pp. 291-305, March 2002.
- [47] R. Rosen, A. Meier, and S. Zandelin, "Energy Use of Set-Top Boxes and Telephony Products in the U.S.," Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory, June 2001.
- [48] K. Roth, F. Goldstein, and J. Kleinman, "Energy Consumption by Office Telecommunications Equipment in Commercial Buildings Volume 1: Energy Consumption Baseline," Arthur D. Little Reference No. 72895-00, January 2002.
- [49] T. Simunic, L. Benini, P. Glynn, and G. De Micheli, "Dynamic Power Management For Portable Systems," *Proceedings of the Sixth International Conference on Mobile Computing and Networking (MOBICOM)*, pp. 11-19, August 2000.
- [50] S. Singh (PI), "NeTS-NR: Strategies to Save Energy in the Internet," *NSF Award Abstract - # 0435328*, September 15, 2004. URL: <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0435328>.
- [51] V. Soteriou and L. Peh, "Dynamic Power Management for Power Optimization of Interconnection Networks Using On/Off Links," *Proceedings of the 11th Symposium on High Performance Interconnects*, pp. 15-20, August 2003.

- [52] M. Srivastava, A. Chandrakasan, and R. Brodersen, "Predictive System Shutdown and Other Architectural Techniques for Energy Efficient Programmable Computation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 4, No. 1, March 1996.
- [53] S. Thomas, and C. Barthel, "www.internet.co2? GHG Emission Trends of the Internet in Germany," Presentation at the IEA Workshop, February 2002.
- [54] "U.S. Nuclear Reactors – Crystal River," Energy Information Administration, Department of Energy, 2004. URL: http://www.eia.doe.gov/cneaf/nuclear/page/at_a_glance/reactors/crystal.html.
- [55] "VERDIEM – Energy Efficiency for PC Networks," 2003. URL: <http://www.verdiem.com/company.shtml>.
- [56] "Wake up to Wake-on-LAN," IBM Corporation, 1996. URL: <http://www.networking.ibm.com/eji/ejiwake.html>.
- [57] "Welcome to the UPnP Forum!," 2004. URL: <http://www.upnp.org/>.
- [58] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, "Self-Similarity through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, pp. 71-86, February 1997.