TCP Window Based DVFS for Low Power Network Controller SoC

Eyal-Itzhak Nave and Ran Ginosar

Department of Electrical Engineering,
Technion--Israel Institute of Technology,
Haifa 32000, Israel
eyal.nave@intel.com, ran@ee.technion.ac.il

Abstract. The decision to enable a network controller to operate at a high performance mode, at the cost of high power, should not rely solely on the amount of data that needs to be transmitted, but also on the ability of the network to deliver it. This work presents a power reduction approach for network controllers using the TCP protocol's unique capability to sense congested networks. Simulations show that it consistently saves at least 10% more energy than work-load only based DVFS throughout various traffic loads and that it nearly doubles the energy saved at various network congestion levels.

Keywords: DVFS, Low Power, TCP, LAN, Network Controller.

1 Introduction

A 2006 study [1] estimates that, in the U.S. alone, annual energy consumption of networked systems approaches 150 TWh, with an associated cost of around 15 billion dollars. The prevalence of networked mobile devices demands longer battery life and less heat dissipation. Data centers growth struggles with the challenges of cooling data center and lowering electricity costs.

In this study we have developed a novel approach, TCP Window DVFS (TWD), for power saving in the most popular computer networks, those using the TCP protocol. Our network DVFS approach, in contrast with previous approaches, determines the DVFS power mode not only according to the work-load (as may be indicated by accumulated packets in buffers). Rather, we also consider the ability to successfully transfer packets through the network. We use the TCP window size to sense network congestion. That window grows with received acknowledgements and is reduced upon packet loss. We compare this method with a simpler DVFS approach, Packet Buffer DVFS (PBD) [2]. Simulation results show that TWD's energy savings are significantly greater than those of PBD, though both lead to major savings in power consumption.

The rest of this paper is organized as follows: Section 2 surveys previous related work. Section 3 describes the proposed "TCP Window DVFS" (TWD). Section 4 describes the simulation that was used to compare energy savings results. Sections 5

J.L. Ayala, D. Shang, and A. Yakovlev (Eds.): PATMOS 2012, LNCS 7606, pp. 83–92, 2013.

and 6 compare energy consumption and saving of TWD across various traffic loads and congestion levels, respectively. Section 7 summarizes this work and offers conclusions.

2 Related Work

In this section we survey previous research aiming to reduce power and energy in networks by modifying protocols of various layers of the OSI model, such as the Data Link Layer (Ethernet) and Transport Layer (TCP).

2.1 Energy Efficient Ethernet

The Energy Efficient Ethernet (EEE) standard (IEEE Std 802.3az-2010) defines mechanisms to stop transmission when there is no data to send. Low Power Idle (LPI) is used instead of the continuous IDLE signal when there is no data to transmit. LPI defines long periods over which no signal is transmitted and short periods when a signal is transmitted to refresh the receiver state to align it with current conditions. [3] shows how packet coalescing can be used to improve the energy efficiency of EEE. EEE is limited to wired network systems using IEEE 802.3 Ethernet protocol. In contrast, our PBD and TWD methods are bounded to neither wired networks nor a specific Data Link Layer protocol, so they can also be applied, for example, to wireless networks. PBD can be utilized in any network where packet buffers are used to store packets before processing. TWD requires, in addition, the usage of TCP as the Transport Layer protocol. EEE is also limited to either LPI mode or full work mode, while PBD and TWD enable multiple DVFS modes for finer tuning of power.

2.2 TCP Level Power Reduction

'Green TCP/IP' has been developed as part of the Energy Efficient Internet Project [4]. It addresses loss of TCP sessions when the CPU is shutting down. Idle hosts are often left fully powered because network protocols and mechanisms fail when the host is not able to conduct basic state-keeping operations. The solution is based on adding a new option in the TCP header ("TCP_SLEEP"), instructing the server to bypass all internal TCP/IP instructions which would drop the connection for that client. Thus, the TCP connection stays open without a need for any activity from the client side (the CPU can shut down). Once the CPU resumes, it can continue sending packets on the open TCP connection without the costly need to reinitiate the TCP connection. However, that solution suffers from three major disadvantages:

1. Energy saving is only achieved when the client is completely shut-down, missing the cases of active idle or low network utilization periods, which comprise a significant part of network activity. Measurements show that the average utilization of desktop Ethernet links is in the range of 1%-5% [5], [6]. Both PBD and TWD provide major energy saving for these low network utilization periods, while maintaining high performance for high utilization bursts/periods.

2. This solution is not adaptive to network conditions. In fact, the "sleep mode" is triggered by a CPU shutdown of the client, regardless of the TCP connection load or network conditions (congestion, link breaks, etc.). Our method is network oriented and adjusts power/performance tradeoff according to TCP connection load and network conditions, taking advantage of data existing in the TCP window.
3. Energy saving is only achieved at the client side. The server side continues to consume energy as if the TCP connections were active in-order to maintain the connections open when clients wake up. In contrast, our solution allows both ends of the TCP connection to use low power mode when possible, thereby enabling double energy saving. Our simulation experiments follow the mutual effects of dynamic power mode changes on both ends of a full duplex TCP connection. Each side includes both RX and TX with independent power managements, comprising a network system with mutual four independent power management systems.

2.3 Power Reduction in Data-Centers and Wide-Area Networks

Data centers are a major source of network energy consumption. The ElasticTree [5], a network-wide power manager, dynamically adjusts the set of active network elements, links and switches, to satisfy changing data center traffic loads. The links and switches that are not needed are turned off. DVFS, as used in our work, enables higher resolution of power management which is not limited to shut-down of a component, but also enables interchanging different work modes.

Chabarek *et al.* [7] use mixed integer programming to optimize router power in a wide area network, by choosing the chassis and line-card configuration to best meet the expected demand. Mandviwalla *et al.* [8] explored using DVS in multi-processor based line-cards. Nedevschi *et al.* [9] investigated network savings with both DFS and DVFS in addition to putting network components to sleep. They propose shaping the traffic into small bursts at edge routers to facilitate putting routers to sleep. Their work compares sleeping vs. rate adaptation in terms of the energy savings achieved across a range of network utilizations. In our work, unlike [7][8][9], DVFS is based not only on the traffic utilization/load, but also on the network congestion and availability.

3 TCP Window Based DVFS

The "TCP window based DVFS" (TWD) policy is based on a simple TCP concept: acknowledgments of transmitted packets indicates that the packets have arrived at their destination and thus the network is able to cope with the current traffic. Failure to receive an acknowledgement results in a sharp decrease of the size of the TCP window because the network may be too congested to successfully deliver the packets of the full window.

Packet Buffer DVFS (PBD) policy [2], as opposed to TWD, determines its power mode based solely on the amount of work to be done, i.e. the size of the packet buffer. When the packet buffer is filled above a threshold, PBD uses high power mode. High power mode maximizes the transmission rate of packets, even when the network is

too congested to enable successful delivery of these packets. Such a policy may result in many lost packets which need to be retransmitted. These lost packets cause a decrease in the TCP window's size, which would limit the number of packets transmitted in parallel and would not allow new packets to be transmitted until the transmitted ones are acknowledged.

In such a PBD scenario the LAN controller is in high power mode, but the actual transmission rate is low, limited by the decreased TCP window. Therefore, there is no advantage in using high power mode when the network is congested, and power is wasted. High power mode should only be used when high performance is useful, i.e., when the network is not congested

Fortunately, the same TCP window, which limits the number of transmitted packets when packets start getting lost, can be utilized to sense congestion in the network. With TWD, power mode is determined by both the packet buffer size and the TCP window size. The packet buffer size threshold cannot be ignored, as high power mode is useless and wasteful when there are only a few packets in the packet buffer or it is empty. Lost packets during network congestion cause the TCP window size to decrease. TWD senses this decrease and affects transition to low power mode. Thus the two indicators, packet buffer size and TCP window size, complement each other in low power LAN controller.

4 Simulation Modeling

To compare energy consumption of TWD, PBD and existing baseline non-DVFS network systems, a configurable DVFS simulation environment was developed, simulating different traffic patterns transmitted during a TCP session with different network conditions. We assume separate DVFS work-points (High/Low) and transmission rates for TX and RX. We further assume that the switching time between power states is negligible, but do take into account the energy overhead consumed for the switching. The simulated system is schematically described in Fig. 1.

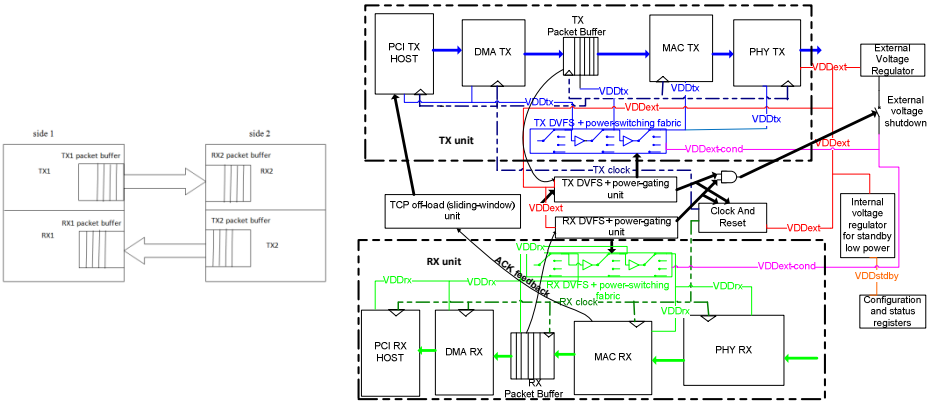

Fig. 1. (left) simulation architecture scheme: two sides, one port each, separate domains for RX/TX per side; (right) functional clocks and voltage domains of one side

In Fig. 2, simulation plots show TWD transitions to low power mode when sensing network congestion as indicated by a sharp decrease of the TCP window size, as opposed to PBD which stays in high power mode during congestion periods. In addition, TWD transitions to high power mode only when the TCP window size reaches the high threshold, while PBD only requires the packet buffer size to cross its high threshold.

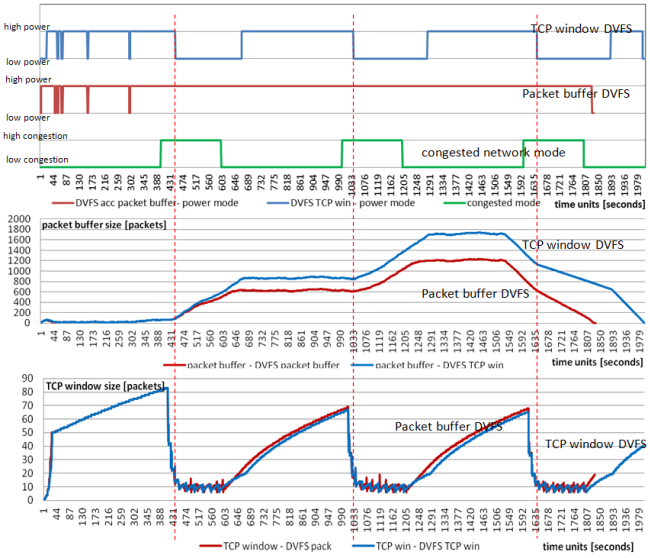

Fig. 2. TCP window DVFS transitions to low power mode when sensing congestion

All simulation runs start with empty packet buffers, and employ a real trace [10] that provides packet arrivals during 1600 seconds. The simulation proceeds beyond 1600 seconds until all packets are delivered and acknowledged.

In the simulation run of Fig. 2, 8,752 and 8,558 packets arrive at the TX of side 1 and side 2 respectively, during 1,600 seconds. The completion time of all these packets and their respective acknowledgements, including lost packets during congested periods, varies across DVFS policy, sides and unit type (TX/RX). The longest simulation run is 2,005 seconds. The network enters a congestion period 400 seconds after the end of the previous congestion period and stays congested for 200 seconds, in which time a packet is lost every 50 seconds.

The total energy consumed by a unit during the simulation run is:

$$E_{unit}=P_{low}*\Delta T_{low} + P_{high}*\Delta T_{high} + E_{lh}*N_{lh} + E_{hl}*N_{hl} \quad (1)$$

where P_{low} and P_{high} are the power consumption of low and high DVFS modes, respectively, ΔT_{low} and ΔT_{high} are the time the unit has operated in low and high DVFS modes, respectively, and E_{lh} and N_{lh} (E_{hl} and N_{hl}) are the energy overhead and number of transitions from low to high (high to low) DVFS mode, respectively. The total

energy of a simulation run is the sum of the energy of all 4 units: TX1, RX1, TX2 and RX2. The energy savings in Sections 5 and 6 are calculated as follows:

$$E_{s(\text{DVFSmodel1 vs. DVFSmodel2})} = E_{total(\text{DVFSmodel2})} - E_{total(\text{DVFSmodel1})} \quad (2)$$

where DVFSmodel1,2 are PBD, TWD or No DVFS as appropriate.

5 Comparison of Energy Saving across Traffic Load Levels

DVFS power mode selection depends on packet arrival rate (i.e. the traffic load). Packet arrival rate is modeled as a Poisson distributed stochastic process. The probability that k packets will arrive in a single time-unit ($\Delta t=1$) is

$$P(k,\lambda)=(\lambda^k e^{-\lambda})/k! \quad (3)$$

where λ is the average number of packets arriving per second. Rather than using the trace in [10], we generate packet arrivals during 1600 seconds according to Eq. (3), for a range of rates. The transmission rate is 3λ and λ at high and low power modes, respectively, because typically doubling the controller frequency from 250 to 500MHZ and raising the supply voltage from 1V to 1.2V would triple the controller processing rate.

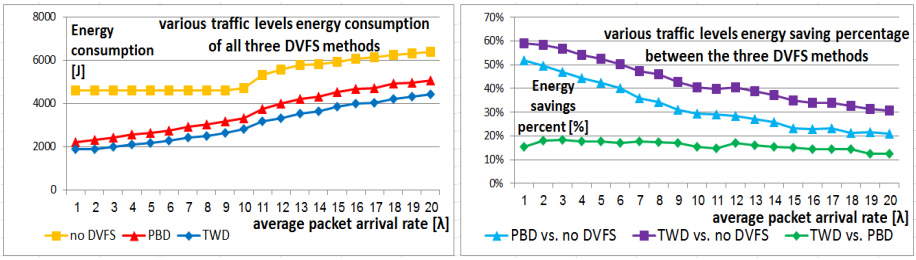

Fig. 3. Different traffic levels: (left) energy consumption; (right) energy saving percentage

Fig. 3 shows the energy consumption and energy saving percentage of the baseline and the two DVFS methods across various packet arrival rates. The network congestion level is held constant for all simulation runs in this section. The time between congested periods is 400 seconds and the length of each congested period is 200 seconds, during which a packet is lost every 50 seconds.

Although packets arrive during 1600 seconds in each simulation run, the completion time varies, as do the time in high power mode vs. time in low power mode and the number of transitions between power modes. According to left graph of Fig. 3, the energy consumption of both PBD and TWD increases with traffic load at an average incremental rate of 158J per unit increase of λ. The maximum energy saving of TWD compared to no DVFS is 2.73KJ, achieved at $\lambda=1$. The maximum energy saving of PBD compared to no DVFS is also at $\lambda=1$ but is about 300J lower: 2.4KJ. Overall, the energy saving of TWD is higher than PBD by an average of 550J. As can be seen from the right graph of Fig. 3, though the decreasing energy saving trends of PBD and

TWD are the same, an approximately constant gap of more than 10% remains throughout all traffic level loads in favor of TWD.

As the traffic load is increased, the energy saving decreases. This is because high traffic load fills the packet buffer above the high threshold causing the network controller to operate at high power mode and increasing the time percentage that the network controller is in high power mode. This sharp decrease of savings becomes more moderate at about λ=10.

The TCP window's size decreases on every lost packet. Fewer lost packets cause less decrease of the TCP window's size, allowing more packets to be transmitted. If the transmitter is in high power mode but the TCP window's size is small, then energy is wasted since the network controller consumes high power but is not able to transmit as many packets as it could have. In addition, when not using DVFS at high traffic load levels ($\lambda \geq 10$), the performance provided by high power mode is insufficient to constantly keep the buffer below the low threshold, as is the case in low traffic load levels ($\lambda < 10$). This results in energy consumption higher than the roughly constant energy consumption at $\lambda < 10$. However, the energy consumption trends of TWD and PBD remain the same as in the low traffic loads. Therefore, a slight increase in energy saving can be observed when $10<\lambda<15$. In these traffic load levels, many packets are lost due to network congestion. The ability of PBD and TWD to transition to low power mode contributes a significant advantage to less energy consumption.

The energy saving achieved with TWD compared to PBD shows (on the left graph of Fig. 3) an increasing trend as traffic load increases. The relative energy saving at the lowest traffic load (λ=1) is nearly doubled at high traffic load (λ=18), rising from 340J to 700J. When TWD energy consumption is compared to that of PBD, the energy saving percentage is roughly stable, ranging from 12% to 18%. In the low traffic load levels the energy saving percentages are slightly higher with energy saving percentages around 18% , while at high traffic load levels ($\lambda \geq 17$) they are slightly lower around 12%-14%.

TWD transits to low power mode as soon as it senses network congestion, providing two means of energy saving: using lower power and reducing the transmission rate to 1/3 of the high power mode transmission rate, resulting with fewer transmitted packets during a packet loss. The effect of these two advantages is more significant when the number of retransmitted packets is higher, in high traffic load levels. This is why the energy saving of TWD when compared to PBD is higher in high traffic loads. However, as apparent from the right graph of Fig. 3, unlike the actual energy saving that reach their maximum value at higher traffic load levels, the maximum percentage of energy saving happens at low traffic load levels, because the total energy consumption of PBD in high traffic loads is higher.

6 Comparison of Energy Saving across Congestion Levels

We now observe the impact of TWD at various levels of network congestion. Ten congestion levels 1-10 were simulated, where 1 is the least congested network level and 10 is the most congested one. A congested network is characterized by lost packets. The more congested the network is, the more packets are lost. A network is

usually not congested 100% of the time. We define simulated congestion levels according to both the frequency and length of the congestion periods and the frequency of packet loss in a congested period. At congestion level 1, the simulation stays 400 seconds in non-congestion mode, and then enters congestion mode and stays there for 200 seconds. In congestion mode, a packet is lost every 200 seconds (once per congestion time period, rather than 4 as in sections 4-5). At congestion level *i*, both the time between congestion periods and the time between lost packets during a congested period are divided by *i*, while the length of the congestion period is multiplied by *i*. As opposed to Section 5, the packet arrival distribution is the same for all simulation runs (extracted from the same real traces [10] as in Section 4). Thus, in this section we isolate the effect of network congestion on the energy saving of each DVFS policy.

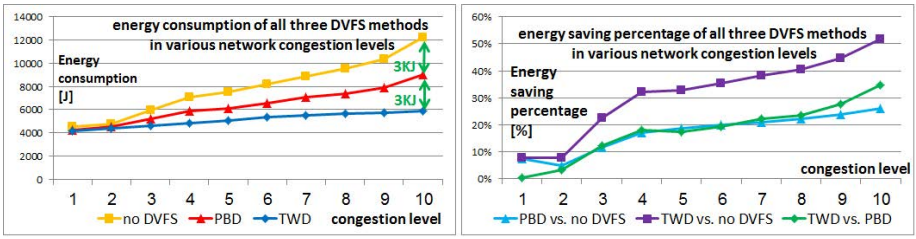

Fig. 4. Different congestion levels: (left) energy consumption; (right) energy saving percentage

According to Fig. 4, at low congestion levels, the energy consumption difference among DVFS modes is small. Level 1 is less severe than congestion in previous chapters. As congestion increases it can be observed that the energy consumption of PBD is about the midpoint between the energy consumption of the baseline and TWD. The energy saving benefit increases with congestion, from less than 1KJ in low congestion levels to more than 3KJ in high levels, in addition to the energy saved with PBD. Clearly, PBD copes well with congestion, and TWD provides even better energy saving. The right graph of Fig. 4 shows that TWD doubles the energy saving of PBD throughout all congestion levels (the blue and green curves are close to each other).

TWD boosts the energy saving percentage in highly congested networks (from 8% to 52%). This is the major benefit from exploiting TCP's ability to sense congested network and react accordingly by transiting to low power mode. Therefore, the TCP window is a better indicator for power mode selection than the packet buffer. As expected, the incremental energy saving in TWD (green graph) increases with congestion. The incremental energy saving of TWD over PBD increases from almost nothing at the lowest congestion level to 35% (3.15KJ). This clearly points out the advantage of using TWD over PBD, especially in high congestion networks.

7 Conclusions

This paper presents a novel approach to power reduction in network controller SoCs, using the advantage of the unique congestion-avoidance feature of TCP to improve previous work-load based DVFS mechanisms. The key idea behind TCP window

based DVFS is that the present work-load should not be the only factor for the decision whether to use high-power/high-performance mode or not. Rather, the ability to efficiently carry-out this work must also be considered. For a network controller, the ability to transmit a load of packets depends on the network congestion level. The novelty of this work is in utilizing the TCP window in addition to the packet buffer load for DVFS decision. We have simulated a network to predict the energy savings of this approach over DVFS based only on the size of the packet buffer, and arrived at the following conclusions:

1. TCP window based DVFS achieves higher energy saving than packet buffer based DVFS thanks to its ability to sense network congestion.
2. Both methods of DVFS reach their peak energy saving in low traffic loads. This is important because the main problem with network power reduction is during idle and low traffic periods.
3. The more the network is congested, the more efficient are both DVFS methods.
4. The advantage of TWD compared to PBD in various traffic loads depends on whether the metric is the magnitude of energy saved or the percentage of the energy saved. The magnitude is higher in high traffic load levels because using low power mode during periods of network congestion avoids many energy-wasteful packet retransmissions and packet loss. On the other hand, TWD achieves higher percentage of energy savings compared to PBD in low traffic loads because of the higher total energy consumption consumed at high traffic load levels.
5. TWD achieves higher energy savings compared to PBD in highly congested networks because of its ability to sense the congestion (via the TCP window).

Table 1. Energy saving percentage results summary

DVFS type	Across traffic loads			Across congestion levels		
	min	max	median	min	max	median
PBD vs. No DVFS	21%	52%	29%	7.4%	26%	19%
TWD vs. No DVFS	31%	60%	40%	7.7%	52%	34%
TWD vs. PBD	12%	18%	16%	0.3%	35%	19%

Table 1 summarizes the minimum, maximum and median energy saving percentage results achieved across various traffic loads and various network congestion levels. Comparing row 2 to row 1 proves that TWD indeed provides more energy saving in every criterion and even doubles the max percentage across congestion levels.

Our simulation model uses only two DVFS power modes in-order to simplify the analysis, enabling a clear picture of the benefits of TWD over PBD. Future work may use more complex models, having more power levels and/or usage of Adaptive Voltage and Frequency Scaling (AVFS) to reflect the activity dependency over time.

References

1. Nordman, B.: Networks, Energy, and Energy Efficiency. In: Cisco Green Research Symposium (March 2008)

2. Nave, E., Ginosar, R.: PBD: Packet Buffer DVFS. Technical report (2012), http://webee.technion.ac.il/~ran/papers/NaveGinosarPacketBufferDVFS2012.pdf
3. Christensen, K., Reviriego, P., et al.: IEEE 802.3az: The Road to Energy Efficient Ethernet. IEEE Commun. Mag. (2010)
4. Irish, L., Christensen, K.: A "green TCP/IP" to reduce electricity consumed by computers. In: IEEE Southeastcon, Orlando, FL, pp. 302–305 (April 1998)
5. Heller, B., Seetharaman, S., Mahadevan, P., Yiakoumis, Y., Sharma, P., Banerjee, S., McKeown, N.: Elastictree: Saving energy in data center networks. In: Proceedings of the 7th USENIX Symposium on Networked System Design and Implementation (NSDI), pp. 249–264. ACM (2010)
6. Christensen, K., Gunaratne, C., Nordman, B., George, A.: The Next Frontier for Communications Networks: Power Management. Computer Comm. 27(18), 1758–1770 (2004)
7. Chabarek, J., Sommers, J., Barford, P., Estan, C., Tsiang, D., Wright, S.: Power awareness in network design and routing. In: Proc. IEEE INFOCOM (2008)
8. Mandviwalla, M., Tzeng, N.-F.: Energy-Efficient Scheme for Multiprocessor-Based Router Linecards. In: Proceedings of the International Symposium on Applications on Internet, January 23-27, pp. 156–163 (2006), doi:10.1109/SAINT.2006.29
9. Nedevschi, S., Popa, L., Iannaccone, G., Ratnasamy, S., Wetherall, D.: Reducing network energy consumption via sleeping and rate-adaptation. In: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, San Francisco, California, April 16-18, pp. 323–336 (2008)
10. TIPC-over-TCP_disc-publ-inventory_sim-withd.pcap, SampleCaptures - The Wireshark Wiki, http://wiki.wireshark.org/SampleCaptures