An Assessment of Power-Load Proportionality in Network Systems

Marco Ricca
Politecnico di Torino
Turin, Italy
marco.ricca@polito.it

Andrea Francini,
Alcatel-Lucent Bell Laboratories
Mooresville, NC, USA
andrea.francini@alcatel-lucent.com

Steven Fortune, Thierry Klein
Alcatel-Lucent Bell Laboratories
Murray Hill, NJ, USA
{steven.fortune,thierry.klein}@alcatel-lucent.com

***Abstract*— Power-load proportionality is a necessary feature in networks that aim at maximizing energy efficiency. Even in networks that handle near-capacity loads very efficiently, energy savings increase substantially if the power consumption closely follows the offered load. Rate adaptation is a common denomination for a set of technologies that operate at different timescales to establish power-load proportionality. To propel their deployment, it is useful to assess how compatible they are with existing network systems and identify the design upgrades that can maximize their energy savings in future networks. The formulation of accurate energy profiles for current equipment is a first step in this direction. We run extensive power measurement experiments to compile the energy profiles of five network systems from multiple vendors. Our results show only negligible signs of power-load proportionality in all five cases: to really make a dent in the carbon footprint and operational cost of packet networks, future system designs must pervasively deploy rate adaptation technologies, especially those that control power state transitions at the packet timescale.**

I. INTRODUCTION

In packet networks, the term rate adaptation designates a broad set of methods aimed at establishing a direct relationship between sustained workload and energy consumption. In an ideal framework for energy efficiency, the network design is optimized to minimize energy consumption under full-load traffic conditions [1]. Rate adaptation additionally ensures that the energy-workload function is linear and that the network consumes no energy when there are no packets to transport [2]. To support such behavior, rate adaptation schemes provide the network systems that they control with a discrete set of operating states, where each state maps a fixed traffic processing rate onto a respective power consumption level. The set may also include a low-power sleep state that suspends traffic processing. The scope of a rate adaptation scheme can range from large portions of a network [3], [4], [5] to individual sections of a single traffic processing chip [6].

Marco Ricca was with Alcatel-Lucent Bell Laboratories when he contributed to this work. This work was supported by the ECONET (low Energy Consumption NETworks) project, co-funded by the European Commission under the 7th Framework Programme (FP7), Grant Agreement no. 258454, and by the U.S. Department of Energy (DOE), award no. DE-EE0002887 (any opinions, findings, conclusions and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the DOE). This research used resources of the ESnet Advanced Network Initiative (ANI) Testbed, which is supported by the Office of Science of the U.S. DOE under contract DE-AC02-05CH11231, funded through the American Recovery and Reinvestment Act of 2009.

Here we classify rate adaptation techniques based on their timescale of operation, which is defined by the switching time needed to transition between states. Demand-timescale rate adaptation (DTRA) techniques control the state of network links and nodes based on expected or measured trends in traffic demands between network endpoints [3], [4], [5]. DTRA state transitions involve network signaling and system-level power cycles, so their timescale ranges from seconds to minutes. Packet-timescale rate adaptation (PTRA) techniques adjust the clock frequency and supply voltage of data-path hardware components to locally maintained workload indicators such as queue lengths and traffic arrival rates [7], [8], [9]. The timescale of PTRA state transitions ranges from microseconds to milliseconds depending on the underlying integrated circuit technology. Bit-timescale rate adaptation (BTRA) also applies to data-path hardware components. Compared to PTRA, BTRA transitions are faster to execute and save less power because they only involve control of the system clock.

To assess the degree of power-load proportionality that the different rate adaptation techniques can enforce, we conduct power measurement experiments on a set of network systems from multiple vendors. In the case of network-wide DTRA techniques, the energy profiles that result from the measurements quantify the benefits of enabling and disabling network ports and possibly also entire line cards and systems based on expected traffic demands; in the case of PTRA and BTRA techniques, the profiles identify the energy-saving margins that are available for the introduction of rate-adaptive hardware components.

The energy profile of a network element maps system and traffic configurations onto power consumption levels, typically by means of a simplified linear model. Examples of system configuration variables that make up an energy profile include the number of cards plugged into the chassis (in slotted systems), the number of ports that exchange traffic over network links, and the transmission capacity provisioned for those ports. Traffic configuration variables include the traffic arrival rate at each network port and the statistical distribution of packet sizes and packet inter-arrival times at ports where traffic is present.

Over the last few years, the accuracy of models and measurement methodologies for profiling the energy consumption of network systems has improved substantially [10], [11], [12], [13], [14]. In our experiments we adopt a model and a measurement methodology that have many similarities with those presented in [13]. While the lower

978-3-901882-56-2 ©IFIP, 2013

degree of sophistication of our test equipment forces us to simplify parts of the model, for example by removing the term for storage energy, the results in [13] indicate that the impact of our simplifications can safely be considered negligible.

The contribution of this paper is twofold. First, we show that even with rather rudimentary test equipment it is possible to isolate crisply the potential for energy saving that is associated with rate adaptation techniques at all timescales. Second, from the observation of the energy profiles that we derive from experimental measurements on five different systems, we conclude that power-load proportionality is poorly supported in commercial equipment.

The paper is organized as follows. In Section II we overview instances of rate adaptation techniques. Sections III and IV describe the systems under test and the power measurement testbed. Section V illustrates our ideal and practical models for energy profiling. We present the results of our measurements in Section VI and draw our conclusions in Section VII.

II. Rate Adaptation Overview

A. Packet-Timescale Rate Adaptation (PTRA)

PTRA techniques target the design of individual hardware components in the data path of network systems. They provide those components with multiple operating states, each state being characterized by a traffic processing rate (expressed in bits per second or packets per second depending on the function of the component) and a corresponding power consumption level. The goal is to minimize the energy spent by the hardware component to sustain the traffic workload that it receives from the data path. State-setting decisions occur at the micro/millisecond timescale, in response to fluctuations in traffic arrival rates and packet queue occupancies.

The authors of [7] studied the application of sleep-state-exploitation (SSE) and rate-scaling (RS) techniques to the links of a network. With SSE a link alternates between only a full-capacity state (at full power) and a low-power sleep state. With RS a link can choose from a set of operating states that lie along a convex curve in the power-rate plane. The specification of the two techniques was refined in [8] with robust constraints on the packet delay degradation that they induce and by formalization of a new hybrid rate adaptation (HRA) scheme that combines the best properties of the two approaches. (We note that several papers from the literature [7],[15],[16] use rate adaptation to designate only rate scaling techniques that do not provide a sleep state. In this paper we follow the convention established in [8], where rate adaptation is a superset for RS, SSE, and HRA techniques.)

A fundamental property of PTRA, not always fully appreciated, is that a mandate to keep the state transition time well within the sub-millisecond range guarantees that the technology is virtually transparent to the operation of the network [9]. If the state transitions took longer to execute, the technology would simply not be suitable for widespread deployment in packet networks. Therefore, provided that the state transition time mandate is satisfied, network links and nodes are never seen missing by the rest of the network, even when most of their hardware components are in their low-power sleep states. Likewise, PTRA is never directly the cause of packet losses or of disruptive degradations in the performance of network protocols and applications.

B. Bit-Timescale Rate Adaptation (BTRA)

BTRA gates the clock signal to eliminate the power consumption associated with bit-level state transitions. Clock gating immediately suspends traffic processing in all portions of a device where it takes effect. A single clock cycle is sufficient to complete the transitions to and from the gated state, so they have no impact on packet delay. However, compared to the sleep state of PTRA the gated state of BTRA does not reduce the dominant component of power consumption, which results from leakage currents, and therefore its energy savings remain marginal.

C. Demand-Timescale Rate Adaptation (DTRA)

An excellent example for the illustration of the goals and mechanics of DTRA techniques can be found in [3]. The paper uses simulations to estimate the energy savings that can be obtained in the Ethernet switching infrastructure of a data center by turning off unused switches, disabling unused ports, and adapting link capacities. The input to the simulation experiments is a 5-day trace of traffic demands averaged over 10-minute periods. A first round of tests produces ideal results under the assumption that a centralized power controller knows ahead of time the evolution of the traffic demands. More realistic results are subsequently obtained with predictors based on real-time traffic measurements. Load prediction errors translate into link overload conditions with higher queueing delays and packet losses, or simply wasted energy.

The portion of the network topology that is subject to DTRA control consists of a set of 1-redundant trees with two tiers of switches. The algorithm that assigns processing jobs to the servers at the leaves of the trees is designed to minimize the overall energy consumption of the two tiers of Ethernet switches. The energy profiles of the switches provide the foundation of the job assignment algorithm. They are based on the following definition of total power consumption S:

$$S = C_0 + \sum_{i=1}^{N_L} L_{0,i} + \sum_{j=1}^{N_P} P_j, \qquad (1)$$

where C_0 is the power consumed by the chassis when idle; N_L is the number of line cards plugged into the chassis; $L_{0,i}$ is the power consumed by line card i when idle; N_P is the number of ports that are connected and enabled; and P_j is the power consumed by port j when enabled, irrespective of the traffic that flows through it.

The best of the three job assignment algorithms studied in the paper yields energy savings up to 75% within the two switched tiers if tangible impairments are accepted with respect to queueing delay and service availability. With a more conservative scheme that avoids any degradation of data center performance the maximum savings amount to 20%.

The authors of [4] apply power-aware routing to variously meshed topologies for IP autonomous system (AS) networks. Compared to the tiered switching networks of [3], the AS networks present different hop counts for alternate paths between endpoints. As a consequence, the energy benefits of any diversion from the basic shortest-path routing are partially reduced by the associated increase in the average number of hops per end-to-end path. The reference model for power consumption only focuses on network ports and excludes contributions from the chassis and line cards:

$$S = \sum_{j=1}^{N_P} \left(P_{0,j} + P_{b,j}\, \beta_j \right). \quad (2)$$

Differently than P_j in (1), the value of $P_{0,j}$ in (2) is obtained when port j is enabled but idle. $P_{b,j}$ is the power consumed by the same port when loaded at full bit rate, which depends on the type of the port and on its rate configuration, and β_j is the bit-rate load sustained by the port ($0 \le \beta_j \le 1$, with $\beta_j = 1$ when the port is fully loaded). The paper evaluates the joint effects of DTRA (instantiated as power-aware routing) and PTRA (only applied to network links, not entire nodes), concluding that power-aware routing is most beneficial when PTRA is scarcely deployed, as is the case in network equipment available today. If PTRA is completely absent but individual links can be turned on and off, the energy savings in one sample topology range between 25% (at 90% of the maximum load) and 50% (at 10% load).

Power-aware routing is applied in [5] to an experimental core network where the continuous transit of packets forces individual nodes to remain powered on without interruption. A mixed integer program that handles binary and continuous variables uses the following linear model to control the distribution of traffic over the network links, switching off unused links and saving additional energy with PTRA in partially utilized links and nodes:

$$S = C_0 + C_b\, \beta_C + \sum_{j=1}^{N_P} \left(P_{0,j} + P_{b,j}\, \beta_j \right). \quad (3)$$

In (3), C_0 is the power consumed by the chassis when idle, C_b is the additional power consumed by the chassis when fully utilized, and β_C is the chassis bit-rate load. The paper avoids a parametric analysis of the optimal solution by assigning a fixed value to every parameter. The model of (3) introduces a chassis contribution that quantitatively dominates over the port terms, with substantial impact on the energy saving metrics network-wide. However, it does not include terms for the explicit contributions of individual line cards, which appear instead in (1). In absence of PTRA, DTRA saves only 0.2% of the overall energy, despite a 34% reduction in the energy consumed by the network links. With ideal PTRA, which scales power linearly with the load in the links and chassis of every node, PTRA alone saves 96% of the total energy, while DTRA only adds an extra 0.1%. The model of (3) rightly takes into account the full-duplex nature of network links and ports. As a consequence, a network port is fully loaded when its traffic load is 100% in both the input direction (from the network to the port) and the output direction (from the port to the network). Accordingly, the port load variable β_j ranges between 0 and 2.

III. Systems Under Test

We obtain energy profiles for five systems under test (SUT's), manufactured by multiple vendors:

ES1—Ethernet switch in fixed system configuration with integrated control and switch module (no slots for plug-in cards), twenty-four 1GbE Ethernet ports (SFP), two 10GbE Ethernet ports (SFP), and AC power supply. The switch supports VLAN and MPLS tunneling for E-Line, E-LAN, and VPLS applications.

ES2—Ethernet switch with twenty-four integrated 1GbE ports (RJ-45), four of which are dual-mode ports that also offer the alternative of loading an SFP module, two 10GbE Ethernet ports (SFP), and AC power supply. The aggregate capacity and functional capabilities of ES2 are the same as those of ES1. One important difference that is worth noting is that ES2 has twenty-four integrated 1GbE ports, whereas all 1GbE ports of ES1 are SFP-ready.

IR1—Edge/aggregation router in fixed system configuration with integrated control and switch module, twenty 1GbE Ethernet ports (SFP), six 10GbE ports (SFP), and AC power supply.

IR2—Aggregation router in fixed system configuration with integrated control and switch module, six $10/100 Mbps$ Ethernet ports (RJ-45), two 1GbE ports (SFP), and DC power supply.

IR3—Aggregation router in modular system configuration with 8-slot chassis. In the IR3 instance available for our experiments, the chassis is populated with one fan card, two control and switch module (CSM) cards, and two 8-port Ethernet adapter cards (EAC's). Each EAC includes six Ethernet ports (RJ-45) and two 1GbE ports (SFP). IR3 also works with a DC power supply.

Due to budget limitations, high-end systems for edge and core routing are not in our set of SUT's. However, the variety of functions, switching architectures, and manufacturers in the set is broad enough to give us good confidence that the energy profiles that we obtain from our measurements constitute a reliable representation of the energy profiles of a majority of the systems that are available today on the market.

Note: Ethernet (RJ-45) identifies an integrated 10BASE-T, 100BASE-TX, or 1000BASE-T Ethernet port. Ethernet (SFP) identifies an Ethernet port that accommodates a small form-factor pluggable (SFP) transceiver. The SFP itself can be of different types depending on the type of cable connector that it supports: 1000BASE-LX and 1000BASE-SX SFP's support optics cables, 1000BASE-T SFP's support copper cables with RJ-45 connectors, and 10GBASE-LW/LR SFP's support optics cables. 10GBASE-LW/LR modules are commonly referred to as XFP's, but throughout this paper we call them SFP's for simplicity of notation.

IV. Power Measurement Testbed

We list the definitions and conventions that we follow in the presentation of our results and describe the equipment that makes up our experimental testbed, highlighting the constraints that it imposes on the execution of the power measurements.

A. Definitions

We refer to an SFP-ready SUT port as *loaded* if it has an SFP module attached; otherwise we call it an *empty* port. We refer to a loaded port or to an integrated RJ-45 port as *connected* if a network cable connects the port to a peering interface on the same system or on a traffic generator/sink, and as *disconnected* otherwise. We refer to a network port as *enabled* if it is configured for operation at a set rate, and as *disabled* otherwise. In general, a port can be switched between the enabled and disabled states when it is empty, loaded but disconnected, and connected. However, we are only interested in the distinction between the enabled and disabled states in the particular case where the port is connected, because this is the kind of state transition that is controlled by DTRA techniques.

B. Testbed Equipment

Fig. 1 shows a schematic drawing of the laboratory testbed where we execute the power measurements. The testbed includes the items listed in the following subsections.

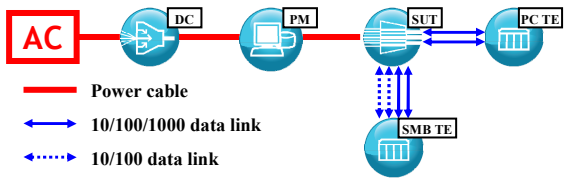

Fig. 1. Experimental testbed for power measurements, inclusive of AC power supply (AC), optional DC power supply (DC), power meter station (PM), system under test (SUT), PC traffic endpoints (PC TE), and Spirent SmartBit SMB-200 traffic endpoints (SMB TE).

1) Power Meter Station

The power meter station (PM in Fig. 1) consists of the power meter and the auxiliary data logging software that runs on a laptop. The power meter is an Extech Instruments 380801 true RMS single-phase power analyzer, placed between the power supply (AC or DC) and the SUT. The meter's resolution is $0.1\,W$ for readings up to $200\,W$ and $1\,W$ for readings between $200\,W$ and $2\,kW$. The data logging laptop acquires power samples at $1\,s$ intervals over the serial port of the power meter. We obtain the $48\,V$ DC power supply for IR2 and IR3 from a Xantrex Technology XKW $1\,kW$ module (DC in Fig. 1). We do not include the power consumption of the DC power supply module in our power measurements. With ES1, ES2, and IR1 the power supply path bypasses the DC module.

2) Traffic Endpoints

For the generation and termination of test traffic we use two desktop computers with 1GbE network cards (PC TE in Fig. 1) and a Spirent SmartBit SMB-200 with 1GbE and $10/100\,Mbps$ Ethernet interfaces (SMB TE in Fig. 1). Each computer runs a Linux OS instance (Ubuntu Release 10.10) and is equipped with one 1GbE network port (RJ-45). We rely on the *iperf* utility for configuration and operation of the traffic sources and sinks on the two PC's. The Spirent SmartBit SMB-200 chassis (firmware version 6.7, umbrella SmartBit release 10.51) hosts two SmartMetrics $10/100\,Mbps$ Ethernet SmartCards (RJ-45) and two GX-1405B 1000BASE-SX Ethernet SmartCards (optics).

We use the two PC's for traffic exchanges with RJ-45 SUT ports at rates up to $1\,Gbps$. We use the $10/100\,Mbps$ Ethernet ports of the SMB-200 for exchanges with RJ-45 SUT ports at rates up to $100\,Mbps$. The two 1GbE ports on the SMB-200 can be used exclusively for exchanges with SUT ports loaded with 1000BASE-SX modules.

3) Network Connectors and Cables

We can rely on two 10GBASE-LW/LR SFP modules for loading the 10GbE ports on ES1, ES2, and IR1. For the SFP-ready 1GbE ports on ES1, IR1, IR2, and IR3 we have two 1000BASE-SX SFP modules and twenty-four 1000BASE-T SFP modules. Notice that, while we have cables for all interfaces, we do not have matching ports on the traffic endpoints for the 10GBASE-LW/LR modules.

V. Energy Profiling Model

In this section we define the linear model that we use for profiling the SUT's of our testbed. We start by showing how we can isolate the contributions of bit-rate and packet-rate loads to the power consumption of a generic system component (chassis, line card, or port). Then we illustrate the model that we consider ideal for unconstrained test environments, where traffic generation resources are unlimited and the power meter has much better sensitivity than our Extech 380801. Finally, we trim the ideal model to meet the restrictions imposed by our testbed.

A. Isolation of Traffic Contributions

In the data path of a network system we can identify hardware devices whose load-proportional component of the power consumption (however small compared to the fixed component) is mostly sensitive to the bit rate of the sustained traffic (e.g., transceivers, switch fabric modules) and others for which it is mostly sensitive to the packet rate (e.g., packet processors, traffic managers). A power meter that only captures fluctuations of the current absorbed by the system cannot directly identify the power consumed by each device, but can detect the effects of varying bit and packet rates on the overall power consumption. We should therefore include independent terms for the bit rate β and the packet rate ρ in the ideal expression of the power consumed by each controllable system component (chassis, line card, and port). However, since β and ρ are not independent of one another ($\rho = \beta / \bar{\sigma}$, where $\bar{\sigma}$ is the *average packet size* measured in bits), the term for packet-rate sensitivity cannot be a function of the packet rate ρ. Instead, the term is a linear function of the average packet

size $\bar{\sigma}$, such that the packet rate contribution is null when the average packet size is maximum (i.e., the packet rate is minimum for the given bit rate), and maximum when the packet size is minimum (the packet rate is maximum for the given bit rate). The following equation expresses our first-order approximation X_t^d of the contribution of each traffic direction d to the power consumed by a generic system component x (chassis, line card, or port):

$$X_t^d(\beta, \bar{\sigma}) = X_b^d \, \beta_x^d \left[1 + X_r^d \left(\sigma_{\max, x} - \bar{\sigma}_x \right) / (\sigma_{\max, x} - \sigma_{\min, x}) \right]. \quad (4)$$

In (4), X_b^d is the *bit-rate sensitivity* of component x in direction d (X becomes C when x is the chassis, L when x is a line card, and P when x is a port); X_r^d is the *packet-size sensitivity* for the same component and direction; β_x^d is the sustained *bit-rate load* ($0 \le \beta_x^d \le 1$); $\sigma_{\max, x}$ is the maximum size of a data packet in the component (e.g., $\sigma_{\max, x} = 1518\ B$ when x is an Ethernet port); and $\sigma_{\min, x}$ is the minimum size ($\sigma_{\min, x} = 64\ B$ when x is an Ethernet port).

B. Complete Linear Model

The linear model that we consider ideal for application in rate adaptation contexts is one that captures the power contributions of all system components whose state can be controlled by external action, whether by network signaling, by system management interface, or by physically plugging or unplugging hardware. These system components include the chassis, the line cards (when present), and the network ports with respective accessories (e.g., SFP modules in our set of SUT's). For every component, there should be one term that expresses the fixed cost of keeping it powered on and one that is sensitive to traffic. The contributions of bit rate and packet rate should be distinguished in parts of the system where the packet size is variable.

The following equation reflects the above requirements:

$$S = C(\beta_C) + \sum_{i=1}^{N_L} L_i(\beta_i^{in}, \beta_i^{out}, \bar{\sigma}_i) + \sum_{j=1}^{N_P} P_j(\beta_j^{in}, \beta_j^{out}, \bar{\sigma}_j), \quad (5)$$

where:

$$C(\beta_C) = C_0 + C_b \, \beta_C \quad (6)$$

is the power consumed by the chassis, inclusive of a fixed term C_0 and a variable term that depends on the aggregate traffic load sustained by the switch fabric;

$$L_i(\beta_i^{in}, \beta_i^{out}, \bar{\sigma}_i) = L_{0,i} + L_{b,i}^{in} \, \beta_i^{in} \cdot \left(1 + L_{r,i}^{in} \, q(\bar{\sigma}_i) \right) + L_{b,i}^{out} \, \beta_i^{out} \cdot \left(1 + L_{r,i}^{out} \, q(\bar{\sigma}_i) \right) \quad (7)$$

is the power consumed by line card i, inclusive of a fixed term $L_{0,i}$ and variable terms that depend on input and output loads;

$$q(\bar{\sigma}_i) = \left(\sigma_{\max, i} - \bar{\sigma}_i \right) / (\sigma_{\max, i} - \sigma_{\min, i})$$

is the *packet-size load*, completely independent of the bit-rate loads β_i^{in} and β_i^{out}; and

$$P_j(\beta_j^{in}, \beta_j^{out}, \bar{\sigma}_j) = P_{0,j} + P_{b,j}^{in} \, \beta_j^{in} \cdot \left(1 + P_{r,j}^{in} \, q(\bar{\sigma}_j) \right) + P_{b,j}^{out} \, \beta_j^{out} \cdot \left(1 + P_{r,j}^{out} \, q(\bar{\sigma}_j) \right) \quad (8)$$

is the power consumed by port j, inclusive of a fixed term $P_{0,j}$ and variable terms that depend on the bit-rate and packet-size loads in the input and output directions of the port. Note again that the packet-size load q is a decreasing function of the average packet size $\bar{\sigma}$: at a given bit rate β, a larger packet size implies a smaller number of packets and less frequent packet processing operations.

C. Discussion of the Complete Model

In this section we illustrate in detail the terms of equations (6)-(8) and refine their definitions where required by the engineering and measurement constraints of our testbed.

1) Chassis Power

The chassis power C is conceivably sensitive to the traffic load, especially if the switch fabric exhibits some degree of modularity with rate adaptation capabilities within each module. In (6) we have no distinct terms for bit-rate and packet-rate contributions because packets typically cross the switch fabric after being segmented into fixed-sized data units with either standard or proprietary formats, setting a constant ratio between the two rates. We do not distinguish between input and output traffic because the amount of packets that enter and exit the central module through the switch fabric interfaces is always the same. The same is not true in individual line cards and network ports, where it is possible to have an unbalance between input and output traffic.

We remark that the type of power meter that we use in our testbed and the absence of rate adaptation capabilities in current-generation switch-fabric hardware make it practically impossible to measure the bit-rate sensitivity C_b of the chassis. More advanced instruments for power measurements, such as those utilized in [13], could isolate the variable terms in the power contributions of the chassis, line cards, and network ports. Instead, in our testbed the sensitivity to traffic shown by the SUT power consumption is so low that it is often masked by the measurement error of the power meter (between $0.05\ W$ and $0.5\ W$). We expect the issue to get gradually resolved in the future, as rate adaptation becomes more pervasive and the necessary instruments become more affordable. For the time being, we consider it acceptable to attribute all traffic sensitivity, including the part that pertains to the cooling fan, to the network ports and reduce the chassis power of (6) to the fixed term alone: $C(\beta_C) \triangleq C_0$.

2) Line Card Power

The line card, when present, is the place where packets that are associated with multiple ports undergo the format conversion from network to switch fabric and vice versa.

It is easy to find a qualitative justification for every term that appears in (7). In the switch fabric adapter, the power contribution of the bit rate dominates over its packet-rate counterpart because the device transmits and receives fixed-size data units. Packet rate dominance over bit rate can be expected instead in the packet processor. However, as already observed for the chassis, the aggregate measurements produced by our power meter compromises our ability to discern the traffic-sensitive power contributions of a line card from those of its ports. As a consequence, we decide again to concentrate all traffic-sensitive terms at the port level, identifying the line card power with its fixed term: $L_i(\beta_i^{in}, \beta_i^{out}, \bar{\sigma}_i) \triangleq L_{0,i}$.

3) Port Power

Due to the simplifications of the two previous subsections, the network port remains the only configurable component of the system where we can retain traffic-sensitive contributions to power consumption. Even the port power model is not exempt from trimming. In fact, because of measurement inaccuracies that are induced by the limited availability of traffic endpoints in our testbed, we cannot differentiate between the values of input and output load parameters. We must resort instead to unified traffic sensitivity parameters $P_{b,j}$ and $P_{r,j}$, and accordingly to unified load variables β_j and $q(\bar{\sigma}_j)$. The value of β_j ranges between 0 (when packet traffic is completely absent) and 1 (when port j sustains 100% bit-rate load simultaneously in both directions).

Preliminary measurements on idle systems show us that the fixed power contribution $P_{0,j}$ of a port j must be split into two distinct terms: the *fixed hardware port power* $P_{0,j}^{(h)}$ and the *fixed software port power* $P_{0,j}^{(s)}$:

$$P_j = P_{0,j}^{(h)} + P_{0,j}^{(s)} + P_{b,j}\,\beta_j \cdot \left(1 + P_{r,j}\,q(\bar{\sigma}_j)\right).$$

$P_{0,j}^{(h)}$ captures the power contribution of port j when it is loaded with an SFP, whether or not the port is enabled for operation. The term obviously disappears in the case of integrated ports. The isolation of $P_{0,j}^{(h)}$ is important because it offers the network operator the option to save energy by unplugging the SFP's of ports that remain disabled for extended periods of time. It also offers system vendors an incentive to add to their designs provisions for controlling this power contribution (and the associated energy waste in the case of disabled ports) via software.

$P_{0,j}^{(s)}$ is the added contribution of a port that is enabled for operation, before it starts handling traffic. Table I lists values for the two terms measured on ES1 with BASE-T and BASE-SX SFP's (configured at 1 *Gbps*), and with BASE-LW/LR SFP's (set at 10 *Gbps*). The switching of individual ports between the enabled and disabled states is one of the primary knobs that DTRA techniques have available for saving energy. Setting the operating rate of an enabled port to a maximum of 10 *Mbps*, 100 *Mbps*, or 1 *Gbps* (and 10 *Gbps* in the case of 10GbE ports) is another dimension of dynamic configuration that DTRA techniques can explore, because each rate generally presents a different value of $P_{0,j}^{(s)}$. In the example of Table I, the measured values of $P_{0,j}^{(s)}$ for a BASE-T SFP are $0.238\,W$, $0.338\,W$, and $1.091\,W$ when the configured rate of operation is 10 *Mbps*, 100 *Mbps*, and 1 *Gbps*.

TABLE I. FIXED PORT POWER TERMS FOR SFP-READY PORTS IN ES1 (BASE-T /SX PORTS SET AT 1 *Gbps*, LW/LR PORTS AT 10 *Gbps*).

	T [*W*]	**SX [*W*]**	**LW/LR [*W*]**
$P_{0,j}^{(h)}$	0.308	0.5	1.2
$P_{0,j}^{(s)}$	1.091	0.3	1.8

4) Simplified Linear Model

The following equation synthesizes the linear model that results from the simplifications of the previous subsections:

$$S = C_0 + \sum_{i=1}^{N_L} L_{0,i} + \sum_{j=1}^{N_P}\left[P_{0,j}^{(h)} + P_{0,j}^{(s)} + P_{b,j}\,\beta_j \cdot \left(1 + P_{r,i}\,q(\bar{\sigma}_j)\right)\right]. \quad (9)$$

We emphasize that the model of (9) derives entirely from simplifications of the model laid out in equations (6)-(8). As the engineering and measurement limitations that warrant the simplifications fade out over time, we expect all the terms of the complete model to gradually reappear in (9).

VI. EXPERIMENTAL RESULTS

In this section we present results from our experiments. We focus on data that gauge the compatibility of existing equipment with DTRA techniques and underscore the need for PTRA support in future system designs.

TABLE II. PARAMETERS OF LINEAR MODEL (1GBE BASE-T PORTS CONFIGURED FOR OPERATION AT 1 *Gbps*)

SUT	$C_0\,[W]$ Chassis, idle	$L_0\,[W]$ Line card, idle	$P_0^{(h)}\,[W]$ Port, fixed hardware	$P_0^{(s)}\,[W]$ Port, fixed software	$P_b\,[W]$ Port, bit-rate sensitivity
ES1 44 *Gbps*	32.4	N/A	0.3	1.1	0.3
ES2 44 *Gbps*	35.0	N/A	N/A	1.0	0.1
IR1 80 *Gbps*	216	N/A	0.2	1.0	1.1
IR2 2.6 *Gbps*	40.4	N/A	0.2	1.0	0.8
IR3 15.6 *Gbps*	54.8	14.5	0.2	1.0	0.8

Table II lists for each SUT the sum of the port capacities (possibly larger than the actual switching capacity) and the

estimated values for five of the six parameters that make up the linear model of (9), in the specific case where the SUT is loaded with BASE-T ports enabled for operation at 1 *Gbps*. The missing parameter is the (port) packet-size sensitivity, whose values are practically impossible to distinguish from zero in the system configurations used in the experiments of Table II. We observe non-negligible values of the parameter only in the case of integrated 10/100 *Mbps* ports in IR2 and IR3 (see Table III for the values measured with ports configured at 100 *Mbps*).

TABLE III. PORT PARAMETERS (10/100BASE-TX PORTS IN IR2 AND IR3 CONFIGURED FOR OPERATION AT 100 *Mbps*)

SUT	$P_0^{(s)}$ [W] Port, fixed software	P_b [W] Port, bit-rate sensitivity	P_r [W] Port, packet-size sensitivity
IR2	0.3	0.1	10
IR3	0.3	0.1	9

TABLE IV. PORT PARAMETERS (1GbE BASE-SX PORTS CONFIGURED FOR OPERATION AT 1 *Gbps*)

SUT	$P_0^{(h)}$ [W] Port, fixed hardware	$P_0^{(s)}$ [W] Port, fixed software	P_b [W] Port, bit-rate sensitivity
ES1	0.5	0.3	0.1
IR2	0.5	0.1	0.8
IR3	0.5	0.2	0.7

TABLE V. PORT PARAMETERS (10GbE BASE-LR/LW PORTS CONFIGURED FOR OPERATION AT 10 *Gbps*)

SUT	$P_0^{(h)}$ [W] Port, fixed hardware	$P_0^{(s)}$ [W] Port, fixed software	P_b [W] Port, bit-rate sensitivity
ES1	1.2	1.8	1.6
ES2	0.9	2.0	0.5
IR1	0.2	1.0	2.9

Table IV lists the parameters of 1GbE ports loaded with BASE-SX SFP's (ES2 is missing because its 1GbE ports are integrated, IR1 because the SMB-200 traffic generator could not be moved to the facility where our instance of the system was located). Table V provides the same information for 10GbE ports loaded with BASE-LR/LW SFP's (10GbE ports are only available in ES1, ES2, and IR1).

The results in Table II indicate that the fixed software port power $P_0^{(s)}$ is by far the dominant port power term in the two Ethernet switches. The traffic-sensitive terms gain relevance in the IP routers, consistently with the increased variety and intensity of the packet processing functions in those systems. Table V shows similar trends for the 10GbE ports, although the traffic-sensitive terms are generally heavier than with 1GbE ports. We note in Tables II and IV the quantitative inversion between fixed hardware power $P_0^{(h)}$ and fixed software power $P_0^{(s)}$ when we replace BASE-T SFP's with BASE-SX SFP's in the 1GbE ports of ES1, IR2, and IR3. Table II also shows that the idle chassis power is much higher in IR1 than in all other SUT's. This is because IR1 is the only system in the set that combines high aggregate switching capacity (80 *Gbps*) with the complex packet processing functions of a router in a non-modular architecture.

We define the margin for saving energy with DTRA techniques as the entire portion of the total energy consumption of a system that is associated with components that DTRA can control. This is clearly a hard upper bound on the amount of energy that DTRA can save. Network topology and traffic demands determine the tightness of the bound in practical applications.

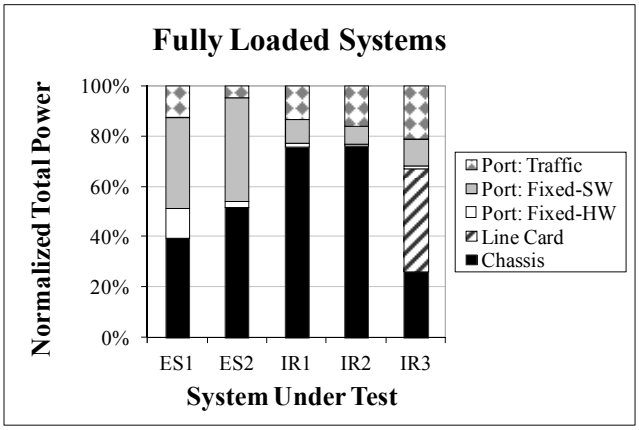

Fig. 2. Estimated breakdown of system power when all ports in the system are fully loaded. The port-traffic quota alone defines the current PTRA margin for saving energy. The port traffic and fixed-software quotas define the DTRA margin.

To quantify the DTRA margin, we must look at the relative weights of the line-card and port terms within the overall power consumption of each system. Fig. 2 shows the breakdown of the total power consumption for the five SUT's when all ports are enabled and fully loaded. In ES1, ES2, and IR1 we configure the 1GbE ports as in Table II (BASE-T at 1 *Gbps*) and the 10GbE ports as in Table V. In IR2 and IR3 the configurations are those of Table II (BASE-T SFP's at 1 *Gbps*) and Table III (integrated BASE-TX at 100 *Mbps*). We normalize the power levels in each column to the total power consumption of the respective system. We obtain the contributions of the port power terms by multiplying the maximum number of ports configurable for each type by the respective per-port values. With IR2 and IR3 the sum of the port capacities in this maximum configuration exceeds by far the actual switching capacity of the system, with the effect of producing overestimated values for the traffic-sensitive power contributions. We would obtain more accurate estimates if we could rely on a larger number of traffic generator ports to pair with the system ports in the power measurement experiments (up to 48 ports with IR3). Still, even if over-estimated and maximized by the assumption of minimum-length Ethernet frames (64 *B*), the traffic-sensitive shares of the total power remain marginal in IR2 and IR3, causing no qualitative impact on the interpretation of the results.

In Fig. 2, the traffic-sensitive terms range between 5% and 21% across the five systems. If we also consider that the two highest values, in IR2 and IR3, are certainly overestimated, the maximum traffic power share is likely well below 15%. We can comfortably conclude that current designs are far from exhibiting the type of rate-proportional power consumption that rate adaptation techniques aim at establishing at the system level. While the indication is disappointing in terms of overall energy efficiency, in light of the results in [4] and [5] it signals that DTRA techniques have a clear window of opportunity in the short term for bringing along important energy savings through relatively simple signaling extensions and software modifications applied to existing hardware platforms.

If we compute the DTRA margin as the sum of the fixed-software and traffic-sensitive port power terms, that is without including the fixed-hardware port power and the line card power, we see that it ranges between 46% and 49% in the two Ethernet switches and between clearly lower values (23% and 32%) in the three IP routers. The potential for DTRA savings increases substantially in modular systems where individual line cards can be switched on and off (+41% in IR3), and even more if DTRA can control the operating state of the entire chassis. However, in network applications that are not necessarily unusual, such as those addressed in [5], it may be likely that an entire system, or even just individual line cards, can never be switched off.

To ensure that the energy savings remain consistently large irrespective of the network topology and application, PTRA capabilities must be pervasively deployed in future generations of hardware platforms. Design challenges and performance properties are well understood for PTRA techniques in linear data-path devices with one input and one output [8], [9]. The same is not true for devices with multiple inputs and outputs like the switch fabric, which typically resides in the system chassis. The challenge for those devices is to achieve direct proportionality between power consumption and aggregate switching throughput irrespective of the traffic load distribution across interfaces. Since the chassis contribution to the total power is always important (between 26% and 76% in the SUT's of our testbed), future research efforts should direct their aim at the identification of viable PTRA solutions for multi-interface devices.

VII. Conclusions

The definition of accurate energy profiles is a critical tool in the process of planning for the short-term (software) and long-term (hardware) design upgrades that can enable better energy efficiency in future generations of network systems. We have defined a linear model that suits well the requirements for supporting the operation of rate adaptation frameworks at multiple timescales. Our model supplies information at the right granularity that DTRA needs for control of the system components that tangibly contribute to the power consumption of individual systems and entire networks. The model also shows that PTRA should be pervasively deployed in future generations of hardware platforms in order to establish true power-load proportionality. Future research efforts should be directed at viable PTRA techniques for data-path hardware components with multiple input and output interfaces.

Acknowledgments

We thank Stephane De Francesco, Larry Friesen, Graeme McClintock, Ramzi Marjaba, Keith Carduck, Rae McLellan, and Adiseshu Hari at Alcatel-Lucent for their contribution to the assembly of our experimental testbed in Murray Hill, NJ. We also thank Brian Terney, Eric Pouyoul, Inder Monga, Michael O'Connor, and Tareq Saif at ESnet for their technical support during execution of our power measurements at the ANI Testbed facility in Brookhaven, NY.

References

[1] R. Bolla., R. Bruschi, C. Lombardo, and D. Suino, "Evaluating the energy awareness of future Internet devices," Proceedings of IEEE HPSR 2011, Cartagena, Spain, July 2011.

[2] L.A. Barroso, and U. Holze, "The case for energy proportional computing," IEEE Computer, December 2007, 33-37.

[3] P, Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "Energy aware network operations," Proceedings of 12th IEEE Global Internet Symposium, Rio de Janeiro, Brazil, April 2009.

[4] S. Antonakopoulos, S. Fortune, and L. Zhang, "Power-aware routing with rate-adaptive network elements," Proceedings of 3rd International Workshop on Green Communications, Miami, FL, December 2010.

[5] A.P. Bianzino, D, Rossi, J.-L. Rougier, C. Chaudet, and F. Larroca, "Energy-aware routing: A reality check," Proc. of 3rd International Workshop on Green Communications, Miami, FL, December 2010.

[6] L. Benini, P. Siegel, and G. De Micheli, "Saving power by synthesizing gated clocks for sequential circuits," IEEE Design and Test of Computers, 11 (4), 1994, 33-41.

[7] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing network energy consumption via sleeping and rate adaptation," Proc. of USENIX NSDI 2008, San Francisco, CA, 2008.

[8] A. Francini, and D. Stiliadis, "Performance bounds of rate-adaptation schemes for energy-efficient routers," Proceedings of IEEE HPSR 2010, Dallas, TX, July 2010.

[9] A. Francini, "Selection of a rate adaptation scheme for network hardware," Proc. of IEEE Infocom 2012, Orlando, FL, March 2012.

[10] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, "Power awareness in network design and routing," Proceedings of IEEE INFOCOM 2008, Phoenix, AZ, April 2008.

[11] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A power benchmarking framework for network devices," Proceedings of the 8th International Networking Conference, Aachen, Germany, May 2009.

[12] O. Tamm, C. Hermsmeyer,, and A.M. Rush, "Eco-sustainable system and network architectures for future transport networks," Bell Labs Technical Journal, 14 (4), February 2010, 311-328.

[13] V. Sivaraman, A. Vishwanath, Z. Zhao, and C. Russel, "Profiling per-packet and per-byte energy consumption in the NetFPGA Gigabit router," Proceedings of IEEE Infocom 2011 Workshop on Green Communications and Networking, Shanghai, China, April 2011.

[14] A. Vishwanath, O. Zhu, R. Ayre, K. Hinton, R.S. Tucker, "Estimating the energy consumption for packet processing, storage and switching in optical-IP routers," Proc. of OFC 2013, Anaheim, CA, March 2013.

[15] S. Ricciardi et al., "Analyzing local strategies for energy efficient networking," Proceedings of Workshop on Sustainable Networking at NETWORKING 2011, Berlin (Germany), 2011.

[16] C. Gunaratne, K. Christensen, and S.W. Suen, "Ethernet adaptive link rate (ALR): Analysis of a buffer threshold policy," Proceedings of IEEE GLOBECOM 2006, San Francisco, CA, November 2006.