◆ Eco-Sustainable System and Network Architectures for Future Transport Networks

Oliver Tamm, Christian Hermsmeyer, and Allen M. Rush

Technology analysis indicates that silicon technology evolution will not be able to catch up with future global Internet traffic growth rates. As a result, current network and system designs as well as network architectures may need to be revised significantly. This paper provides a detailed analysis of various network element types (e.g., IP routers, Ethernet switches, and synchronous optical network/synchronous digital hierarchy [SDH/SONET] switches) and their components with respect to power dissipation © 2010 Alcatel-Lucent.

Introduction

Energy consumption efficiency is gaining greater and greater attention in the future evolution of telecommunication networks. As demand continues to grow for greater network bandwidth, power consumption is becoming a major constraining factor in ongoing network capacity expansion. Future network architecture and systems technology choices will affect the power consumption efficiency of the network overall. As a consequence, it is expected that future network architectures and system technology choices will be heavily influenced by power consumption profiles.

The overall percentage of the Information and Communications Technology (ICT) segment today, estimated at 2 to 2.5 percent of total global energy consumption [14], is projected to grow substantially as the world economy becomes more and more networked and more and more activities go online. Current estimates for year-over-year global Internet traffic growth vary but tend to converge at 40 percent to 50 percent [3, 9].

Advances in silicon integration have enabled network processing capacity and bandwidth gains, but power efficiency of such technology advances has not kept up. As a result, as network capacity growth continues, the energy consumption costs of operating the network are increasing.

To illustrate the problem, at the 2007 Nature Photonics Technology conference in Tokyo, Japan, data was cited claiming IP routers would consume 9 percent of Japan's total electricity production by 2015, and nearly 50 percent in 2020, based on today's technology [4]. Also, at the same conference, Tomonori Aoyama from Keico University, and formerly of NTT's Network Innovation Labs, noted "that by 2020, telecoms would move from switching terabits to petabits, but based on today's technology a 100 Pb/s IP router would consume 10 Megawatts and require a nuclear power station to supply it with electricity" [14].

Although energy prices have stabilized from the dramatic increases of recent years [12], the cost of powering the network, as well as corporate social responsibilities regarding global climate change, are significant motivators driving the need for power efficiency improvements of communications networks. Energy costs account for as much as 2 to 2.5 percent of total telecom operational expenditures (OPEX) globally and this is trending upwards [14].

Bell Labs Technical Journal 14(4), 311–328 (2010) © 2010 Alcatel-Lucent. Published by Wiley Periodicals, Inc.
Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/bltj.20418

Panel 1. Abbreviations, Acronyms, and Terms

ASIC—Application-specific integrated circuit
ASSP—Application-specific standard product
ATM—Asynchronous transfer mode
CAM—Content addressable memory
CDR—Clock and data recovery
CMOS—Complementary metal-oxide semiconductor
CO—Central office
CPU—Central processing unit
DEMUX—Demultiplexing
DiffServ—Differentiated services
DPI—Deep packet inspection
DRAM—Dynamic random access memory
DSLAM—Digital subscriber line access multiplexer
DWDM—Dense wavelength division multiplexing
FEC—Forward error correction
FPGA—Field programmable gate array
GbE—Gigabit Ethernet
GPP—General purpose processor
GRE—Generic routing encapsulation
HO—Higher order
ICT—Information and Communications Technology
IGMP—Internet Group Management Protocol
I/O—Input/output
IP—Internet Protocol
ISPF—Incremental Shortest Path First
IT—Information technology
ITRS—International Technology Roadmap for Semiconductors
LAN—Local area network
LC—Local controllers
LH—Long haul
LO—Lower order
MAC—Medium access control
MEMS—Micro-electro-mechanical systems
MPLS—Multiprotocol Label Switching
MSA—Multiplex section adaptation
MSTP—Multiple Spanning Tree Protocol
MUX—Multiplexing
OA—Optical amplifier
OAM—Operations, administration, and maintenance
ODU—Optical data unit
ODUk—ODU level k
OEO—Optical-electrical-optical
OLT—Optical line terminal
OOO—Optical-to-optical
OPEX—Operational expenditure
OSI—Open Systems Interconnection
OTN—Optical transport network
PDH—Plesiochronous digital hierarchy
PoS—Packet over SDH/SONET
QDR—Quad data rate
QoS—Quality of service
RAM—Random access memory
RSTP—Rapid Spanning Tree Protocol
SDH—Synchronous digital hierarchy
SerDes—Serializer/deserializer
SFP—Small form-factor pluggable
SI—Sideband interfaces
SONET—Synchronous optical network
SRAM—Static RAM
STS1—Synchronous transport signal 1
TCAM—Ternary CAM
TDM—Time division multiplexing
TEEER—Telecommunications Equipment Energy Efficiency Ratings
UPS—Uninterruptible power supplies
VPLS—Virtual private LAN service
VPN—Virtual private network
VPWS—Virtual private wire service
WDM—Wavelength division multiplexing
XFP—10 Gb small form-factor pluggable

As a consequence of going "green," support for environmental initiatives is becoming more and more critical to communication network providers in terms of both demonstrating a positive environmental consciousness to the public and as well as to benefiting their bottom line (i.e., reducing operational expenses). Specific goals for energy consumption reduction can be found in the Corporate Social Responsibility statements of many, if not most, service providers.

BT and Verizon provide two important examples of this trend. BT's corporate social responsibility statement has declared the company's intent to reduce its carbon footprint by 80 percent by 2016 [2]. Network power efficiency is a critical element of reaching that goal along with other areas such as use of sustainable non-greenhouse-gas-emitting power generation (i.e., solar and wind), vehicle fleet fuel efficiency, and paperless billing and office processes. Verizon mandated a 20

percent reduction in power requirements for new network equipment deployed after January 1, 2009 [13]. In addition, Verizon established Telecommunications Equipment Energy Efficiency Ratings (TEEER) as a measurement methodology and requirements for specific categories of network products.

Historically, network equipment technology advances have steadily increased system capacity, port density, throughput, and other parameters without an emphasis on power efficiency. However, the amount of power available to a standard rack or lineup of bays in a central office (CO) has had a static upper limit due to cooling and power distribution infrastructure. Much of the current generation of central office network equipment has pushed system density up to this maximum limit of power per rack at ~15 kW (based on Telcordia Technologies GR-63-CORE 600 × 600 × 2100 mm racks). Nevertheless, additional advances in system capacity are needed to accommodate the increasing bandwidth demands on the network. The feasibility of continuing such growth is questionable given the existing cooling and power distribution practices in today's central offices.

The need for improvement in network equipment power efficiency is also motivated by the additive effects of CO power distribution and cooling. It is estimated that for every 1 watt of power consumed by the network, as much as 2.5 watts is consumed in total when cooling, power distribution, and uninterruptible power supplies (UPS) are taken into account [5]. Although novel methods of removing the vast amount of heat generated by high density network equipment are being investigated, dealing solely with cooling aspects of the energy consumption problem should result in diminishing returns.

The highest returns will result from reducing the power consumption of the network itself while achieving ever-increasing bandwidth and port density.

This paper discusses issues of network power consumption, benchmarks current technologies and product categories, and provides analysis on how network and system architecture choices can affect power efficiency. One important observation made is the relation of protocol layer support in a network node to the power consumption of that node. Network nodes that restrict traffic processing functions to the optical layer (i.e., layer 1) have superior power efficiency over network nodes that are processing traffic at the Internet Protocol (IP) layer (i.e., layer 3).

While the focus here is on the transport, aggregation, and switching portion of communication networks, it is expected that power consumption of the network's customer premises equipment and subscriber access equipment—i.e., digital subscriber line access multiplexers (DSLAMs) and optical line terminals (OLTs)—will exhibit similar trends.

Power Analysis

An analysis of various network element types can validate or invalidate the underlying network paradigms just because of power—a conclusion which is also extremely helpful in response to the public incitements for reduction of CO_2 emissions.

Up to now, power dissipation requirements for telecommunication infrastructure equipment have impacted its architecture and design mainly in the form of engineering constraints within the system boundaries.

For example, hot spots on circuit packs must be avoided by careful airflow and heat-sink design, in order to not overheat temperature-sensitive optical components and semiconductors. Backplanes and connectors must deal with increasingly high supply currents within the shelf.

The purpose of the following analysis is to identify the main contributors to power consumption of data and transmission systems and to contrast various system types by their usage of native technologies, such as optics, general purpose central processing units (CPUs), application-specific integrated circuits (ASICs), field programmable gate arrays (FPGAs), and the like.

While system power analysis has mainly focused on improvements of individual components that are close to the limit of power dissipation, and mechanical and thermal design has been chartered with the objective to cope with worst case equipment configuration, it is now necessary to derive generic trends in order to analyze feasibility of tomorrow's networks.

System Categories

We assumed in the beginning of our analysis that there is a strong relationship between power per given bandwidth of data and the amount of processing on this data. Furthermore, networks are built by a combination of specific system types. Therefore we have started by differentiating network systems by functionality and have put them into the following five categories.

1. IP routers covering:
   - Internet Protocol versions 4 and 6 (IPv4, IPv6) unicast and multicast routing.
   - Layer 2 and Layer 3 virtual private networks (VPNs) and Multiprotocol Label Switching (MPLS).
   - Multi-service supporting SDH/SONET (PoS), PDH, ATM, and Ethernet physical interfaces.
   - Differentiated services (DiffServ) quality of service (QoS) with 8 QoS per port.
   - Layer 2 and layer 3 access control (stateless), metering/marking, statistics/accounting.
   - Large routing tables capable of supporting full Internet routing.
   - Tunneling (MPLS and generic routing encapsulation [GRE]).
2. Ethernet switches covering:
   - Ethernet bridging (802.1d, 802.1q, 802.1ad).
   - Layer 2 VPNs (virtual private wire service [VPWS] and virtual private LAN service [VPLS]).
   - Layer 2 access control, policing/metering/marking, statistics/accounting.
   - Hierarchical QoS.
   - Internet Group Management Protocol (IGMP) snooping, Rapid Spanning Tree Protocol/ Multiple Spanning Tree Protocol (RSTP/ MSTP).
3. SDH/SONET switches:
   - Higher order SDH/SONET mapping/demapping.
   - Overhead processing.
   - HO SDH/SONET connectivity.
   - SDH/SONET OAM.
   - Linear and ring protection.
   - Meshed restoration and optical control plane.
4. Optical transport network (OTN) switches:
   - Higher order optical transport network (OTN) mapping/demapping.
   - Overhead processing.
   - ODUk connectivity.
   - OTN operations, administration, and maintenance (OAM).
   - Meshed restoration and optical control plane.
5. OOO (all optical) switches:
   - Wavelength division multiplexing (WDM) muxing/demuxing.
   - Optical connectivity.
   - Optical amplification.
   - Optical OAM.

These categories and related functionalities do not span the complete set of network functionality and system types, but they cover the major portion in metro and wide area transport systems.

Generic Power Analysis Model/System Differentiation

It is inevitable that with a growing demand for network capacity in the future, network and information technology (IT) power consumption is set to grow at the same scale, unless system vendors redesign their products, and network operators re-architect their networks and change their equipment purchasing criteria.

With the multifold motivators for power reduction and power efficiency improvements being much more severe than just a design constraint to achieve savings in operational expenditure without significantly increasing equipment cost, vendors and operators now need to rethink their strategies in many areas, e.g., the cycle at which legacy equipment is replaced with modern, power efficient systems on the operator side. On the supplier side, the architectures that build tomorrow's networks, the components that are required, and the underlying technology must be gauged in order to validate them along with the expectations of their future power consumption profile.

In order to help vendors and operators to proceed beyond their view on a single "box," a power benchmark is required between various system categories. In the same vein, the power analysis shall reveal the major underlying technologies and architectural paradigms that construct those systems.

When combined with the prevalent expectations of the industry segments that deliver these technologies, an extrapolation is made possible that provides

the power evolution of the different system types. As this is of interest in order to predict the power consumption of future networks in a broader sense, it also allows us to compare the "amps per rack" characteristics between different vendors, and to identify system types that may break another limit, which is the power consumption per rack in the CO.

There are many ways to look at power consumption of (transmission) systems. Very generally speaking, systems consist of a number of components that can be arranged in a flat *bill of material*. While this bill would provide the most accurate view of the power consumption, it is at the same time difficult to interpret, so therefore an abstraction is necessary. A functional view would classify the components on the bill into, e.g., switching, storage, processing, optical termination, and control. Taking boundaries of physical entities into account, a subsystem view can be established which allows separate scaling of functions to some extent, i.e., the ability to grow system transport capacity without changing the control portion of the system. Another view, similar to the bill of material perspective, is a view by technology, like that of complementary metal-oxide semiconductor (CMOS) in its various forms (scale in nm, FPGA versus ASIC, dynamic random access memory [DRAM]), discrete electrical components, electromechanical components, and the like. Finally, a view based on the Open Systems Interconnection (OSI) model for transmission systems results in a layered approach (e.g., layer 1, 2, 2 +, 3 . . .), which provides interesting results for network architecture considerations.

As there are no industry standards and metrics available for a system power analysis, we are taking a three-step approach, starting from a typical transmission system structure, as shown in **Figure 1**.

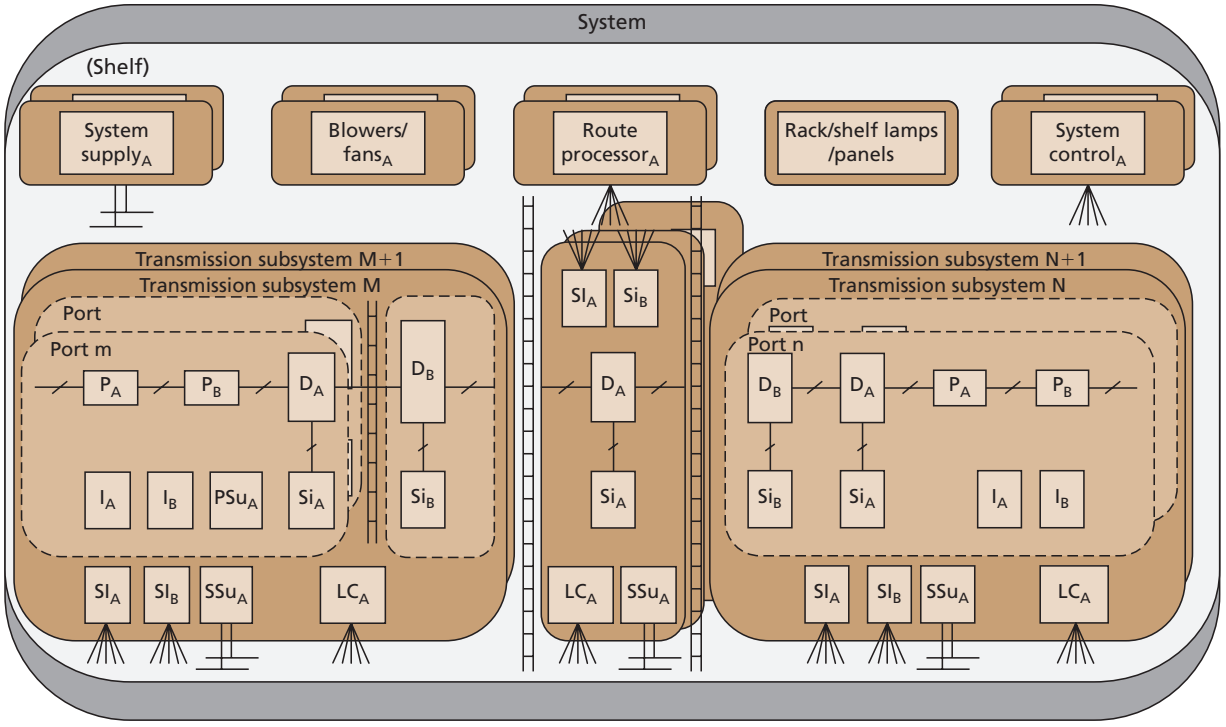

Figure 1.
Transmission system with combined subsystem and functional decomposition.

Figure 1 decomposes a transmission system into centralized resources, like "shelf," "fans," "system control," or "power supply." These functions may occur in a protected or unprotected manner. Centralized switching elements, or subsystems providing connectivity, may also be combined and the variety of arrangements is large. "Transmission subsystems," usually pluggable units that allow exchange of functions, or upgrade of the system, can be further decomposed into "port" functions and "fabric adapters." Furthermore, they may host local controllers (LC), decentralized power supplies (SSu), controller infrastructure, and sideband processing devices and interfaces (SI).

Port functions and fabric adapters can further be segregated into traffic processors ("P", "D") and sideband functions ("Si").

For the purpose of extrapolation of system power figures into the future, the level of detail depicted in Figure 1 is difficult to maintain. Therefore, a more abstract view is created. In the first step we will decompose various types of telecommunication equipment into a set of functional and/or organizational entities. This decomposition shall allow a classification of the parts and subsystems of a specific network element type into comparable categories.

In the second step we will analyze the underlying major technologies or mix of technologies typically present per category. It is important that systems are decomposed in step 1 such that a functional or organization view is maintained, while within every category, various system types may utilize different technology to achieve comparable functionality.

In the third step, we will take a horizontal view across the various categories, and their underlying technologies, in order to identify common technology usage.

Decomposition

System level power analysis is usually done by thermal and mechanical design teams that are asked to improve the airflow architecture and the cooling capacity with ever-growing power dissipation values in a given footprint, with little motivation or charter to reduce power consumption. The change of direction in this analysis is to identify the root causes—the sources drawing power—and reduce the demand for energy at the source. These considerations are rather driven by OPEX analysis, "green" considerations for reduction of the CO_2 footprint, and, especially in the scope of this analysis, by the projection of the future capacity demand against the limits of cooling and power supply in a certain footprint.

Hence, the metric that is applied and the ways in which power is calculated are chosen in a way that allows reports on underlying technology levels. In similar approaches found in the literature [1, 6, 7], power has been allocated by subsystem or component type, or a mixture of both. In [11], the decomposition becomes clearer but is only applied to high-end IP routers. The resulting taxonomy is presented in **Figure 2**. Organized into eight higher level entities,

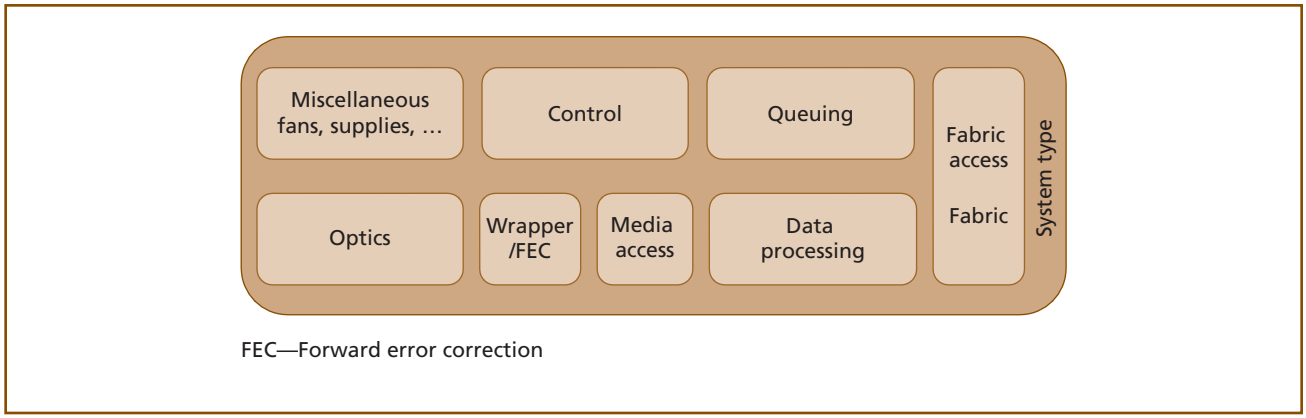

Figure 2.
Abstract decomposition for transmission systems.

any transmission system type in our scope can be decomposed in a way that allows a mapping to these entities. Sometimes, a block coincides with the boundaries of a complete subsystem (e.g., fabric, fabric access). In other cases, functions of a specific subsystem contribute to different categories, or a single entity/category needs to be spread across multiple subsystems, when a comparison of actual power figures is intended.

Examples: SDH/SONET cross connect and router/packet switch. The system architectures of an SDH/SONET cross connect and a large-scale IP router or packet switch are similar. In the aggregation-, metro- and core-backbone networks, which are within the focus of this analysis, those systems are built to utmost density and reliability at lowest cost. Central elements are split from the detachable units that host interfaces of flexible type, speed, and reach. The centralized infrastructure is designed to operate with multiple client protocols, and it is duplicated or made redundant in other arrangements in order to guarantee service delivery. Switch matrices in time division multiplexing (TDM) systems that typically go down to synchronous transport signal 1 (STS1) granular switching are usually fully duplicated in a hot-standby manner, while data systems typically require overspeed in the active switch matrix, and, because duplication is more costly, deploy n:m (all-active) redundancy schemes.

Control functions are needed in order to manage the system itself, to communicate with the outside network (element) management system or with an automatic control plane. IP routers require strong CPU horsepower for operation of routing and signaling protocols, and they come with sophisticated traffic management functions and many functions on the sideband interfaces for deep packet inspection purposes.

The classification of those system types is depicted in **Figure 3**.

Although the blocks that are present in the three system types above (the IP router, packet switch, and SDH/SONET cross connect) are similar, there is differentiation needed in the block analysis. As an example, for an IP router, the "routing and signaling" function is a subsystem that typically requires multiple CPUs with significant amounts of memory. For an SDH/SONET cross connect, this is an important, though only accompanying and less powerful subsystem.

Queuing and traffic management are functions that do not have equivalent entities in an SDH/SONET cross connect system. The detailed differences of the functional entities will be described further in the following sections.

Examples: OTN cross connect systems and all-optical switches. OTN cross connect systems and all-optical switches have fewer functional entities than the systems described previously. Because they operate with signal structures of larger size, they present a significantly smaller number of instances that have to be managed, supervised, and handled by a control plane. In turn, functions for, e.g., routing and signaling are rather lightweight in such architectures.

The class of OTN cross connects still provides for grooming capabilities and for restoration on a per-client basis. To support grooming and aggregation of lower order ODUk signals, electrical components are required for de-multiplexing and switching of the sub rate and full rate signals. Therefore, effort is required for de-framing and de-multiplexing. All-optical switches, which do not convert optical signals to electrical signals, and vice versa, simply switch entire signals with a certain amount of performance monitoring and fault management, found in the "optical OAM" function.

Functional mapping. In this section we will describe the differences of functions that are mapped to each category per network element type. Several vendors offer multiprotocol or hybrid systems, i.e., systems that can operate in either TDM, OTN, or packet switching mode. In most cases, those systems can be operated in the various operational modes concurrently. The efficiency in comparison to other systems depends on the flexibility of the system with respect to the center stage matrix. Sometimes, separate matrices are required for TDM and for packet operation, while other systems have matrices that are oblivious to the data protocol and do not require costly exchange or dual type matrices with additional physical and thermal footprint.

Analyzing a specific product, we can derive separate fact sheets for operation in either TDM or packet mode. As yet, we have not investigated hybrid operational

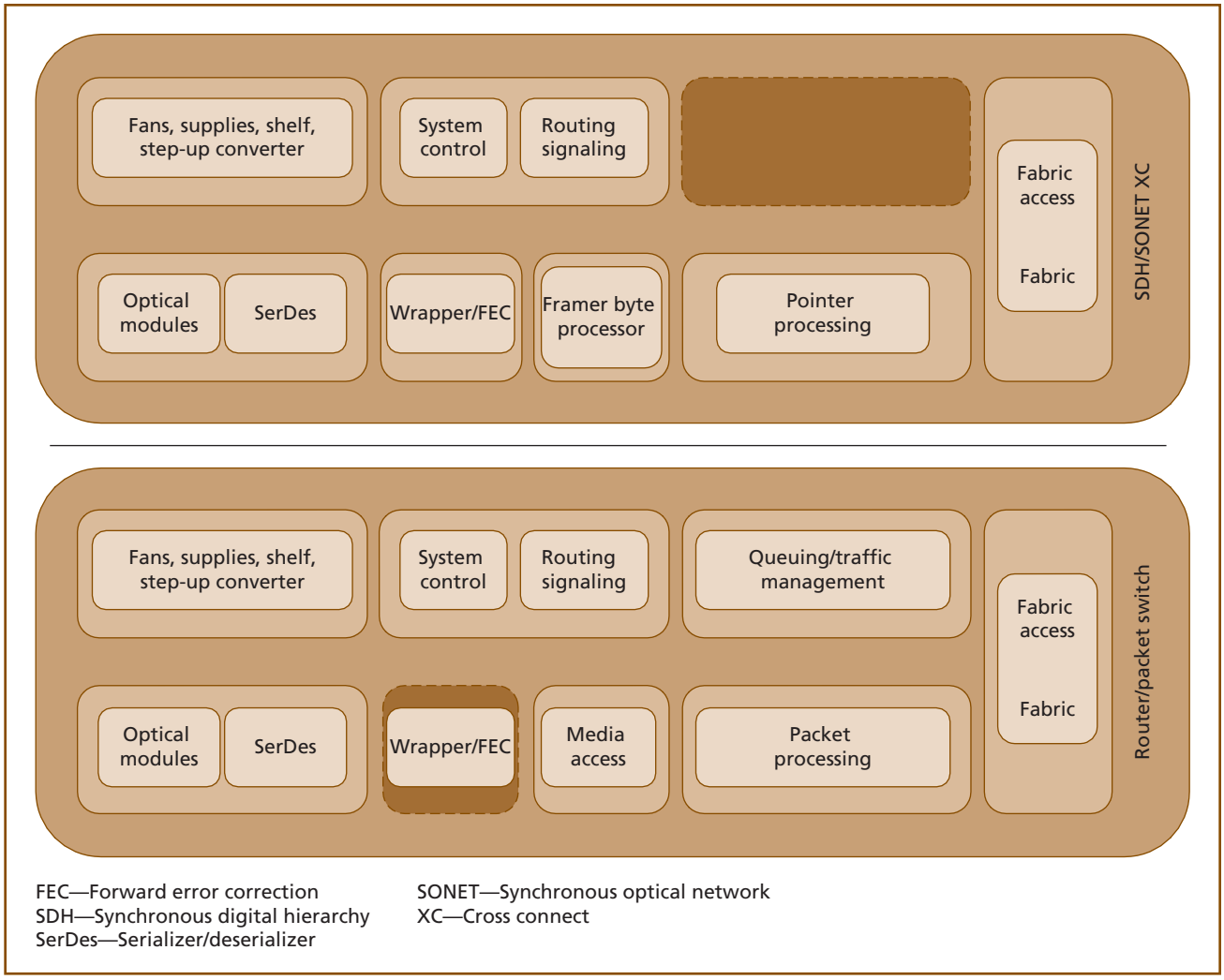

Figure 3.
Classification of routers, packet switches, and SDH cross connect systems.

modes of such systems. Instead, we used those systems in an either-or mode of operation.

Optics. In any system type we have been looking at, we considered the same type of optical module and line termination. Certainly, a WDM long haul system can provide for long reach colored optical modules with high-end performance, but the analysis tries to consider each network element type in a similar, thus comparable network situation. In order to compare the inner parts that make up the system, and to compare the overall power consumption between system types, there should not be a difference in the optical front end. Therefore, the interface type was assumed to be intermediate reach (less then 40 km), uncolored (i.e., black and white) optics, without a dense wavelength division multiplexing (DWDM) front end.

In the system-type benchmark, we considered only the optical module and the serializer/deserializer (SerDes) function, be it a separate device, like an SFP or XFP module, or an integral part of the optical module, as is the case with multiplex section adaptation (MSA)-based optical modules. If no data are provided, we assume ~3.5 watts for a 10 G XFP module, and ~1.2 watts for the SerDes function. The SerDes function is typically built in CMOS-silicon. Depending on the type of optical module, a mixture of optical, discrete, and CMOS or other technologies can be

assumed. The model may also cover optical-electrical-optical (OEO) systems with long haul (LH) DWDM optical front ends by changing the respected power profile of the optical front end.

Wrapper and FEC. This block includes the digital wrapper as is present in today's OTN systems and systems that are equipped with at least an OTN front end termination. Forward error correction (FEC) capabilities are also located within this functional block. Both parts are always built out of CMOS silicon technology, either as separate components or as part of other CMOS entities (e.g., multiplexer devices, SerDes).

Medium access. Electric termination and monitoring of the client are handled in the "medium access" function. Simply, in a packet switch or IP router, this represents the medium access control (MAC), while in an SDH/SONET cross connect, the framer and HO/LO multiplexer, plus fault management and performance monitoring primitives, and, e.g., second generation, are located within this functional entity.

Traffic processing. Traffic processing covers different functions per system type. In a TDM cross connect, pointer processing and adaptation are performed in this task block. In packet switches, packet processing takes place here.

Packet processing includes packet classification, metering, and policing. It also includes sophisticated deep packet inspection (DPI) functions in IP routers. Ternary content addressable memories (TCAMs) are memories specifically used in this context; they are located in the sideband to store, retrieve, and match strings by algorithmic searches. DRAM or static random access memory (SRAM) is used to store large amounts of data (e.g., MAC addresses) in the sideband for similar functions.

Forwarding decisions and internal header processing are other tasks allocated to this block. All of these functions are implemented in CMOS silicon technology; however, they come with different thermal and functional scales. For example, as TCAM is extremely expensive and requires a large budget for thermal dissipation and sophisticated heat sinks, growth can only be expected on a moderate scale. Still, alternative approaches are being investigated at this point in time by the industry to move the functionality of TCAMs into standard DRAM.

Functions above which are not present in OTN cross connects and all-optical switches are not represented in **Figure 4**.

Queuing and traffic management. The only systems that host these types of functions are packet switches and IP routers. In all other evaluations, these functions are left empty. It must be noted that several transmission systems exist that can be operated in either TDM or packet mode. The "queuing and traffic management" functional block will only be considered when a hybrid system is 100 percent equipped as a packet switch.

Queuing and traffic management functions deal with the user traffic but do not account for the management of the fabric itself (which may also require coarse-grained queuing and scheduling, together with matrix arbitration, but these functions are located to the "fabric access" block).

These functions are hosted purely by CMOS-ASICs in the data path, accompanied by packet memory (typically DRAM) and linked-list memory (e.g., quad data rate [QDR]) in the sideband. Performance monitoring, statistics gathering, and pre-processing are further functions of this category that require additional resources (often located in separate components in the sideband).

Fabric access and fabric. This block aggregates the functions used to access, operate, and protect the fabric. Large core transport nodes and aggregation transport nodes are required to provide multi-service functionality as well as flexible switching granularity in support of TDM, packet, and OTN services. Fabric structures can be TDM-based, cell- or packet-based, or combinations of both, depending on the system application. In an all-optical switch, only switching devices are present (e.g., micro-electro-mechanical systems [MEMS]-fabric), and dedicated access components are not necessary. IP routers and packet switches typically require overspeed in support of multicast and QoS guarantees, resulting in larger active matrix arrangements. The access devices perform fabric load balancing and queuing, while arbitration and fabric control are located on a central subsystem, together with a shared memory switch fabric or a different architecture.

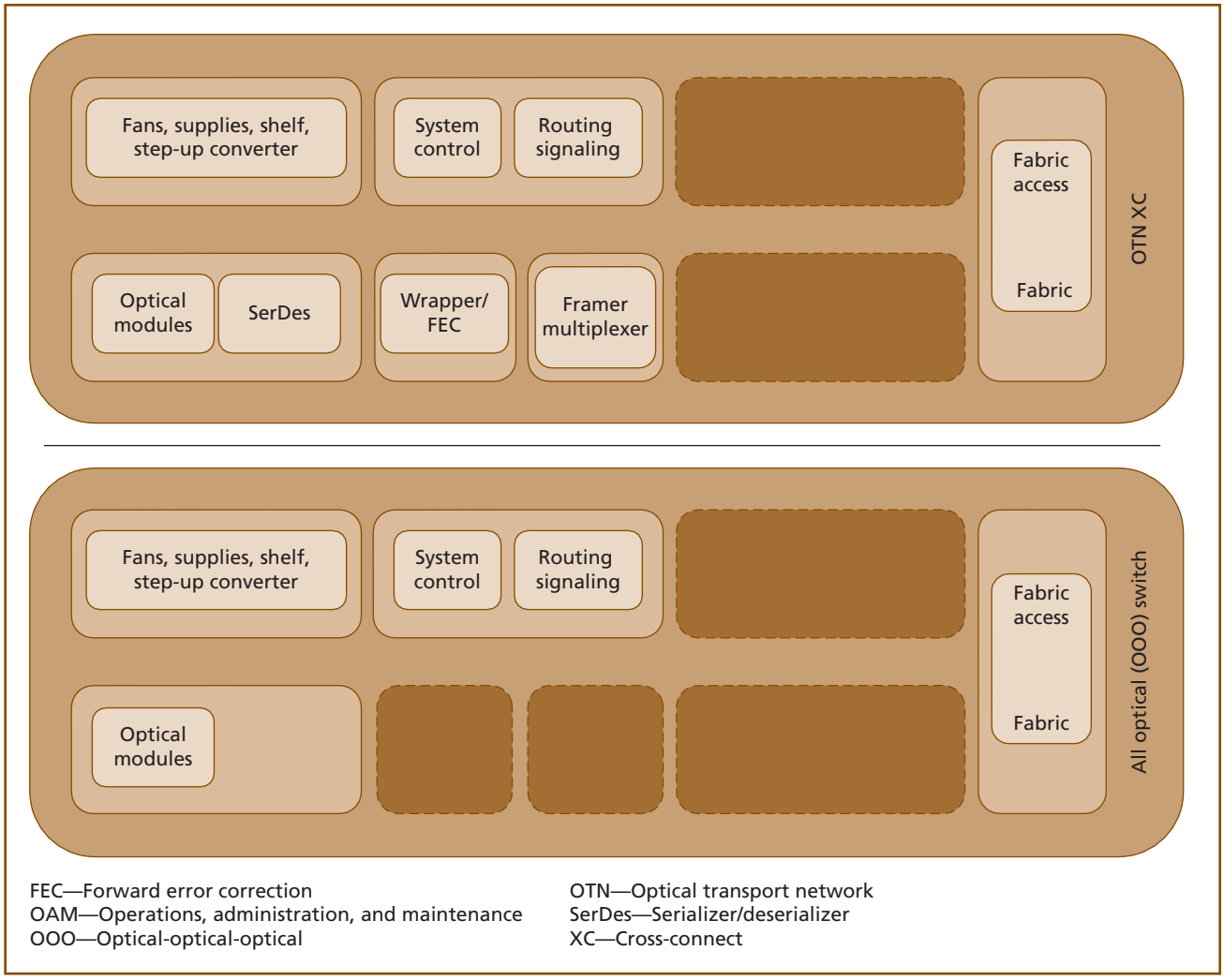

Figure 4.
Classification of OTN cross connect systems and all-optical switches.

Service availability is a key property, implemented in the different system types in different ways. While IP routers and packet switches typically provide for m:n redundancy, TDM and OTN systems are usually 1+1 protected, where the protection resource is in a hot-standby mode.

Control

System control and system management, i.e., an external management interface, command interpreter, and subtending translation functions, are components which are present in every system type. TDM systems (SDH/SONET cross connect systems, OTN systems) usually support a control plane function for automatic setup and teardown of connections through the network. Often, this control plane function is hosted on a dedicated subsystem. In a similar way, packet switches and IP routers require massive CPU power for the reliable operation of dynamic connection setup, and routing- and signaling protocols, like Incremental Shortest Path First (ISPF), Border Gateway Protocol (BGP), or Spanning Tree Protocol, just to name a few. Notably, system controllers and route processors have to be duplicated in order to increase their availability figures.

Deployed technologies in this area encompass general purpose processors (GPP) based on CMOS technology, all kinds of memory types, standard CMOS application-specific standard products (ASSPs), and discrete components for the physical interfaces.

The same holds for distributed entities that control the subsystems. Local controllers, used for device control and configuration, statistics retrieval and correlation, and port-level protocol handling, communicate via backplane interfaces and links beneath each other, or with the central system controllers. In this category, internal timing distribution and collection networks, overhead collection and distribution networks and related processing functions, and other infrastructural functions and their related components and interfaces are taken into account.

Miscellaneous

In this category we collect shelf-level functions and infrastructure. Fan units and blowers are electromagnetic components with a significant power draw. The mechanical sizes of transmission systems have remained more or less constant over time, as they are usually constrained by the footprint in the operators' offices. But power consumption and, as a consequence, thermal dissipation of systems are steadily increasing with system density. Heated air has to be removed from the system via air exchange. This results in a significant amount of fresh air that has to be moved through the system, which can be regarded as an airflow resistor. Therefore, the fan units are expected to provide high speed and high pressure, and they are built redundantly to cope with intermediate failure.

Power supplies, step-up converters, fuses, and the like, are built by components that transmit the aggregate power of all system parts. Power losses in these functions can be regarded as power consumption which scales linearly with the power required for the other system functions, including fans.

Power draw from backplanes, user-panels, rack-top lamps, and other central functions are also considered under the "miscellaneous" category.

Power Characteristics of Exemplary Systems

The following data represents results gathered in an analysis internal to Alcatel-Lucent which scrutinized a variety of systems built by different business groups. Some of the systems we analyzed appeared on the market only recently, while others have been deployed for five years or more. In order to prepare a timely comparison snapshot of the market, we normalized the resulting power figures to the year 2008.

By decomposing the architecture of a system built in 2003, for example, extrapolations in each of the eight categories resulted in an improved overall system power figure for 2008. Basis for the extrapolation was knowledge of technological developments and architectural improvements in every category, blended with assumptions on feature improvements and compromising factors. The latter accounts for the fact that systems are not typically built from scratch, taking, as in a greenfield approach, the best-of-breed components with the latest technology advancements into consideration. Instead, each system has one or many predecessors, and many items evolve into a new generation, such as infrastructural entities, mechanics, and core components, as well as software.

Building a system is a multidimensional optimization, and many compromises have to be made when the architecture is deployed. Up to now, power consumption has been a less dominant consideration. Therefore, the relative power consumption of transmission systems is usually improved implicitly, e.g., as a consequence of higher density transmission components and their newer technology. But the total power values of such systems are still growing with the absolute system size. According to the outlook of this report, this may change dramatically in the near future.

Figure 5 shows a factor of about 6 in total power consumption between IP routers and all-optical switches (based on metropolitan area distances). The power breakdown into the eight categories reveals that three functions dominate IP router power: data processing, queuing, and overhead in switching/fabric access. The latter is due to the over-speed requirements between all fabric inputs and outputs, compared to zero over-speed in all-optical switches and SDH/SONET cross connects. Queuing is a function only present in IP routers and data switches. Data processing is a resource consuming task, as, e.g., pattern-matching and complex algorithmic searches have to be performed in real time at wire speed. Considering 100 Gigabit Ethernet interfaces with a packet rate of up to 150 million packets per second,

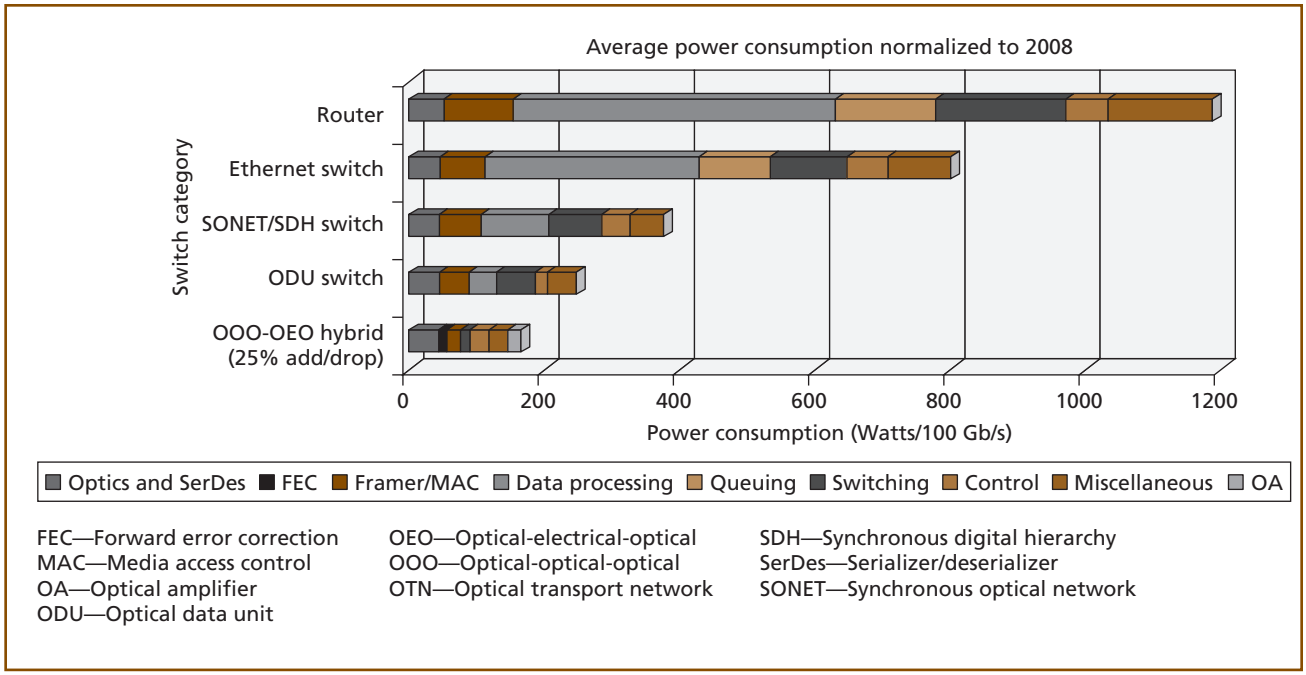

Figure 5.
Power consumption per 100 Gbps for different system types.

wire speed processing results in a packet arrival event every 6.7 ns. Parsing fields of multiple octets against tables with millions of entries requires hardware support, but at the same time the process must take place intelligently in order to leave the system with the right amount of flexibility to support future protocol changes, which are likely to happen in the data world. In turn, those functions (e.g., ternary content addressable memory) are extremely power-hungry, are extremely costly, and require a lot of real estate.

Breakdown of Power Characteristics by Technology

Besides the categories "optics," "miscellaneous," and "optical amplifiers" (OAs), all categories are dominated by CMOS silicon up to 95 percent. The technology is used for ASSPs, FPGAs, GPP CPUs, and custom ASICs.

The most stringent observation from Figure 5 is the fact that data processing, queuing, and switching are doubling/tripling the power required between SDH/SONET cross connects, Ethernet switches, and IP routers. This power increase is a direct result of the CMOS functionality, accompanied of course by the usual losses in supplies, additional control, and an increase in fan power.

Technology Analysis

How the individual technologies applied to telecom equipment may evolve will depend both on principle technology factors like CMOS integration steps as well as on boundaries the system architecture dictates, like overall system power budget together with functional and mechanical decomposition. These together define the footprint per component with respect to functionality, power dissipation, and real estate. For example, consider that the latest CMOS technology and maximum die sizes might allow for much higher integration, but system architecture (i.e., cooling) constraints may not allow leveraging this. This results in constraints to use smaller dies, while other applications (e.g., mainstream CPUs in personal computers) may take advantage of such.

Silicon Technologies Applied in Telecom Equipment

A quick analysis looking for "fat rabbits" in power dissipation indicated that the majority of power dissipation in OEO-based telecom equipment is related to

functionalities being implemented in large-scale CMOS devices. These devices cover the data path from input/output (I/O) front end functions like SerDes, clock and data recovery (CDR), and MAC/framers down to data processing, card interconnect, and switching functions, as well as system and network control functions. This is no surprise, but it is worth saying that even for OEO systems spanning DWDM LH applications, the CMOS-related electrical component is the first to consider for power optimization. With this said, system capacity and systems density for OEO-based switching systems are mainly restricted by CMOS technology limits as mentioned before. Of course, first generations of high-speed optics (like today's 100 G LH optics) consume significant board space, reducing density, but this is mainly due to time-to-market constraints, allowing second generations to shrink significantly and putting the density limits back to the electrical domain.

CMOS Silicon Evolution Technology Aspect

Evolution in CMOS is multifaceted. The pace of evolution depends on technical enhancement in respect to structure pitch (integration level), maximum speed, and power per function. As said, in shelf-based telecom equipment, the power budget per device is relatively low as the options for cooling techniques are more restrictive than, say, a desktop computer, and the environmental conditions are more challenging than in office/home environments. Systems must remain fully operational up to a 50 degree Celsius ambient temperature [10]. On the other hand, telecom equipment has very challenging requirements with respect to integration level and speed to address expected bandwidth growth. Looking at the silicon evolution trends reported in the International Technology Roadmap for Semiconductors (ITRS) [8], the pace of CMOS silicon integration improvements has been and is expected to remain slightly different than those for DRAM, complex logic, and Flash memory applications. An internal analysis of a large set of shelf-based systems (14–18 slots per row, forced air cooling) revealed that the upper power boundary of large scale CMOS devices is typically in the range of 50–70 watts per device package. This is mainly due to the limited cooling options in such slot-based systems. In principle, it would be possible to increase the power per device by using larger and more efficient heat sinks, but this would result in reduced density and less flexibility due to bigger slots and therefore fewer slots per row. Higher airflow may result in exceeding noise emission limits, which are usually already at their boundary. Looking at the power profiles defined by the ITRS, the power profile in telecom equipment is therefore below even the so-called cost performance profile allowing a maximum of ~100 W per large CMOS device [8]. Furthermore, the ITRS model includes a substantial increase in power per device of roughly 50 percent between 2007 and 2016 and beyond. It is unlikely that such a power increase can be accommodated given the mechanical constraints of telecom equipment, and therefore we expect the power profile per device will stay roughly flat. As a consequence, it is expected that the pace of further integration on CMOS devices applied in telecom equipment we have analyzed will be lower than calculated in the ITRS "cost performance profile." It has to be noted that even this cost performance profile logic type is far away from catching up with Moore's law (doubling the amount of transistors every two years per device).

To calculate the integration limit of telecom logic, as shown in **Figure 6**, we've applied an upper device power limit of 70 watts and scaled that with the cost performance profile. As a result the annual density increase (2010–2020) for telecom-related logic is expected to be at ~26 percent while Moore's law is based on ~41 percent density improvement per year. Underlying data for the graph was taken in part from [8].

Increasing the speed of CMOS devices is starting to saturate from 65 nm and 40 nm and further technology steps may provide only very marginal increases in speed. Technology integration (more transistors) will remain the only enabler to address further bandwidth demands. Based on this, we expect a roughly linear dependency between density increase and supported bandwidth in the future. But as the gap between technology evolution and bandwidth growth continues to expand, we expect a severe issue in addressing future growth if no revolutionary silicon technology improvement to increase speed and to lower power materializes. The only alternative

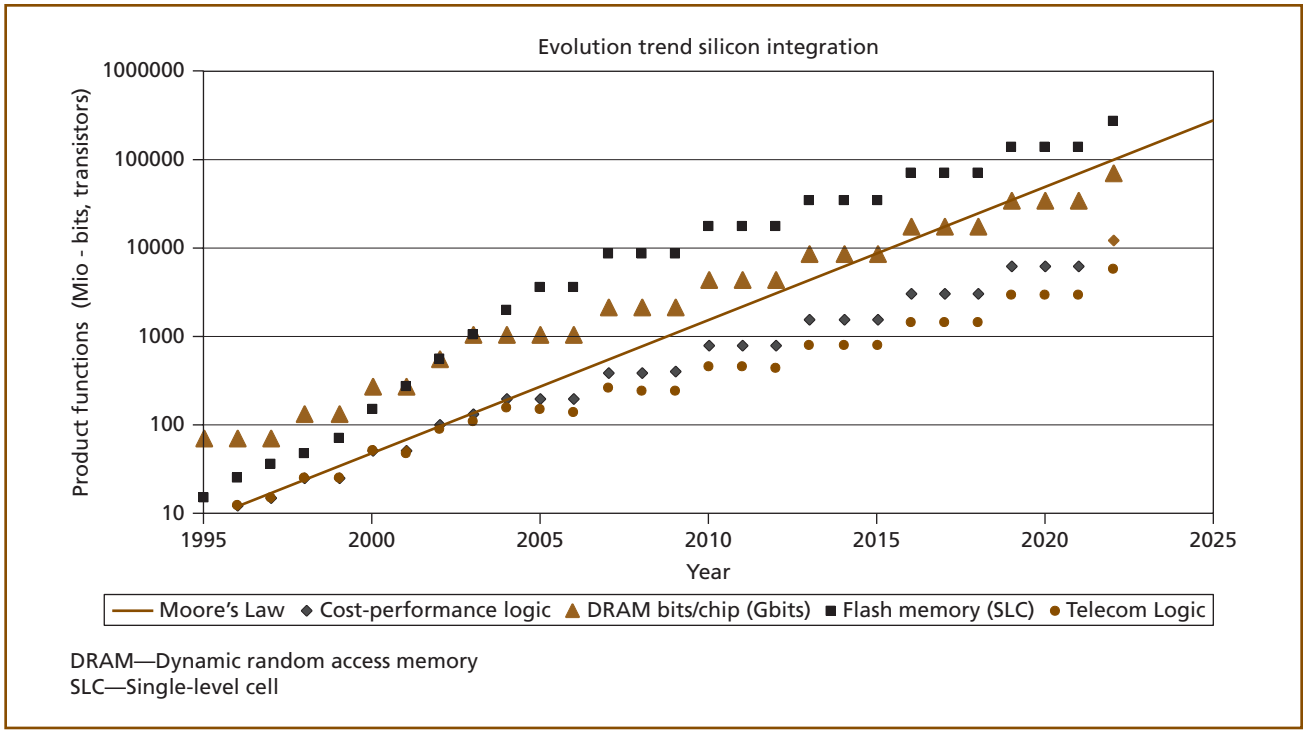

Figure 6.
Roadmap for silicon integration including telecom-applicable logic devices.

appears to be moving the bandwidth to lower networking layers, and thus leveraging their lower power-per-bit profiles.

CMOS Silicon Evolution Commercial Aspects

Our system models include a continuous adaptation/upgrade to the latest available CMOS technologies, but they also take into account that adoption of new technology usually lags a little behind the availability of the newest silicon technology. The model is based on the fact that telecom systems traditionally comprise a mix of technology generations due to the overall complexity and lifetime of these systems. We have applied a sliding window over time using the latest three available technologies (e.g., 90 nm, 65 nm, and 40 nm) to represent a realistic benchmark. Some systems applying only the latest CMOS technology may demonstrate a slightly better power profile, but these systems will also age out, as we do not envision that system manufacturers or network operators can afford to adopt a system lifetime matching the incremental CMOS technology improvement, which is shorter than a two year period.

Optics Technology Evolution

Novel optical technologies like silicon photonics may significantly reduce power and increase density by integrating modulators, lasers, diodes, and waveguides into a single silicon device. Nevertheless, the overall power contribution of these devices is so minor we do not expect a significant impact on the general trend. Nevertheless, optical modules may play an important role in the achievable density due to the high thermal sensitivity of these devices. While standard CMOS devices may operate at ~120 °C to 125 °C die temperatures, optical devices are significant more sensitive and need to be run at significant lower temperatures which may add cooling complexity, resulting in further lower system density.

Power dissipation trend of network element systems. We have applied the CMOS evolution trend results on the various system functions and have put these into a generic power analysis model. This has been performed individually per system category. We also used a mix of three CMOS technology generations per system and applied the sliding window mechanism

over the time scale. Furthermore, the results have been benchmarked on multiple proof points versus existing system implementations in the market to validate the applicability of the model.

Looking at the results, shown in **Figure 7**, we report the following conclusions.

First there is a strong relation between the networking layer the system is operating at and the power profile per capacity. For example, the difference in power dissipation between an IP router function operating at the IP layer and an OTN switch operating on ODU circuits is a factor of ~5. Secondly, the relative power difference between the networking layers remains quite constant over time. This is mainly due to the fact that all system types benefit in the same way from CMOS technology evolutions.

A further surprising result of the system modeling is that the power decrease over time on system level is substantially lower than we would expect based on the ~26 percent annual density increase (2010–2020) of CMOS devices. The primary reason is that power dissipation related to backplane and inter-chip interconnect is not expected to follow the same steep curve as silicon. Due to the fact that gate count improvement is mainly eaten up to address bandwidth growth, a significant vertical integration is not expected, especially on memory-intensive applications like network processing and traffic management. This of course does not exclude vertical integration in the cost of bandwidth per device. Furthermore, attached functions like fans and power supplies stay flat on the power profile.

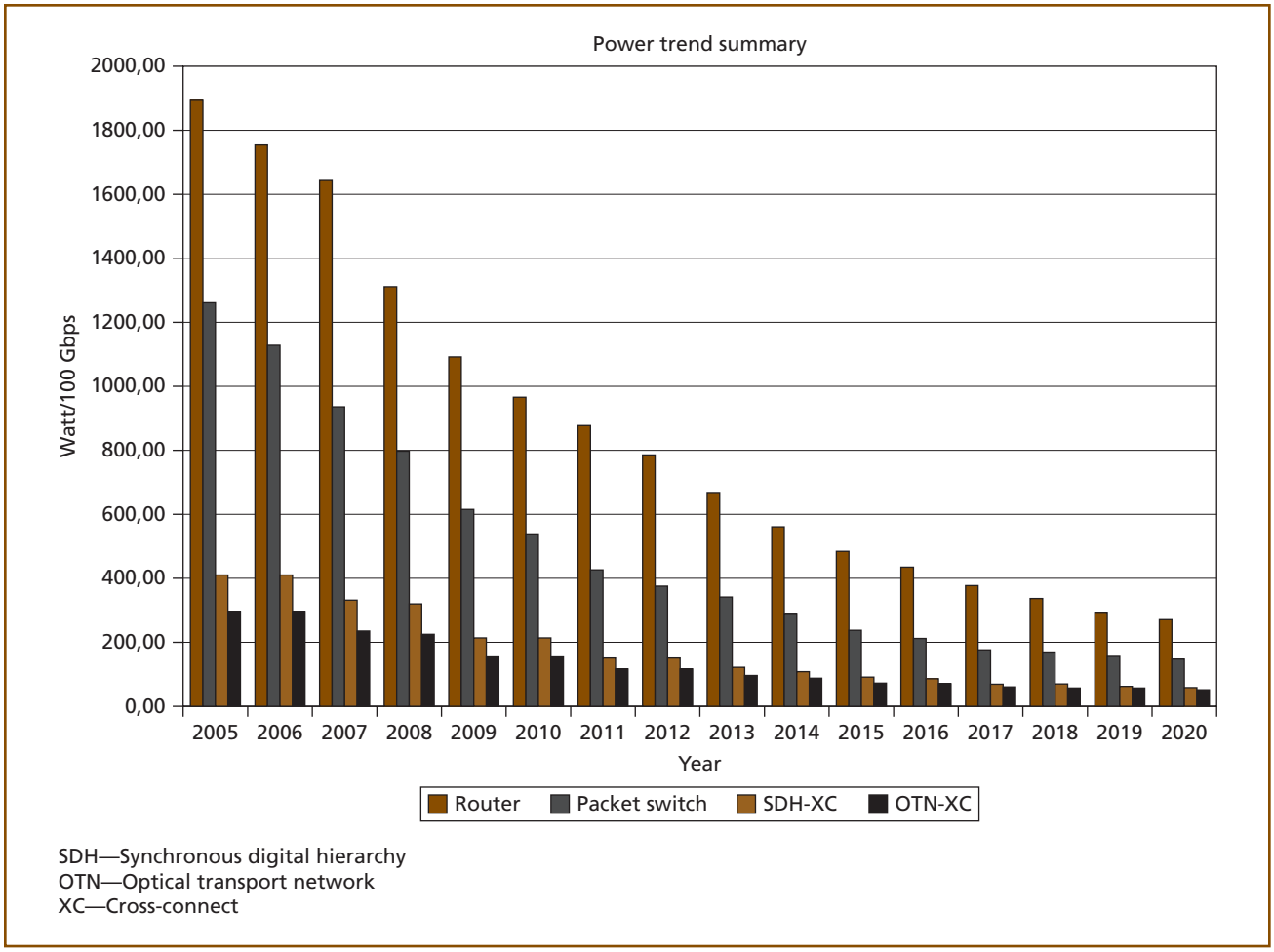

Figure 7.
Power dissipation trend depending on system type.

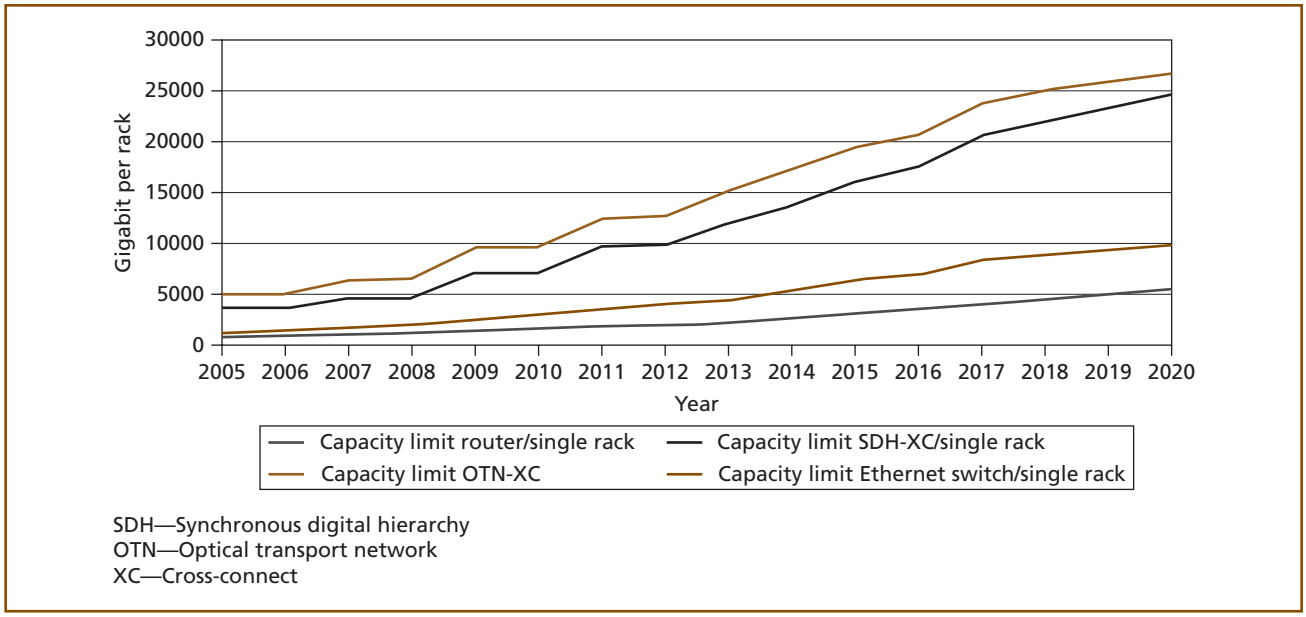

Figure 8.
Rack capacity trend.

System limits. A further interesting aspect is related to system limits. If we apply boundaries for maximum power per rack, it is possible to translate this into theoretic capacity limits as a function of technology evolution. In the following example we have applied a 15 kW upper limit per standard 600 × 600 mm rack.

As expected from our earlier power profile analysis, the results plotted in **Figure 8** show that there is a significant difference in achievable rack density with respect to the network layer. Also expected is that the relative density difference remains nearly constant. As interconnections between shelves are very costly and reduce efficiency, a key network sizing exercise is to avoid these interconnections. So to keep the power and cost profile as low as possible, in OEO systems it is essential that density per shelf is maximized while shelf interconnect is avoided. As OTN and SDH/SONET switches support a significantly higher density than, e.g., IP routers and packet switches, this presents a compelling argument to move bandwidth to lower layer systems.

Conclusion and Outlook

In the analysis we have shown that using a generic model for OEO-based network elements, it is possible to compare and benchmark different network systems. We have looked at CMOS devices as the main contributor to power dissipation in OEO switching systems, and how silicon evolution will evolve and will have impact in density of future systems. As overall global IP bandwidth is expected to grow at a rate between 40 percent and 50 percent, the key conclusion is that silicon evolution alone will not be able to address the expected need for bandwidth growth. It is expected that only a change in network paradigm, e.g., moving as much traffic as possible to the lower layers, may help in addressing future traffic demands. Of course this is not a trivial task as OTN switches cannot easily substitute IP router functionalities, but trying to offload IP routers as much as possible by moving traffic to lower layers might become a key requirement to gain time to sustain the network growth longer. As a next step, we propose to perform detailed network case studies, applying these power profile results, and to reveal the optimal network architecture serving a given application, scalability, and service scenario.

References

[1] S. J. Ben Yoo, "Power Consumption in Optical Packet Switches," Proc. Eur. Conf. on Optical Commun. (ECOC '08) (Brussels, Belg., 2008).

[2] BT Group, Changing World: Sustained Values 2008—Innovation and Implementation, 2008, <http://www.btplc.com/Societyandenvironment/Ourapproach/Sustainabilityreport/pdf/2008/BT_CSR_08_SinglePage.pdf>.
[3] Cisco Systems, "The Exabyte Era," White Paper, Jan. 14, 2008, <http://www.hbtf.org/files/cisco_ExabyteEra.pdf>.
[4] D. Cooperson, J. Mazur, and M. Walker, Increased Focus on Network Power Consumption to Lower OpEx, Go Green, Ovum, Mar. 9, 2009.
[5] Emerson Network Power, "Energy Logic: Reducing Data Center Energy Consumption by Creating Savings That Cascade Across Systems," White Paper, 2009.
[6] G. Epps, "Electronic Routers: State of the Art and Future Perspectives," Proc. Optical Fiber Commun./National Fiber Optic Engineers Conf. (OFC/NFOEC '08) (San Diego, CA, 2008), Workshop on Optical Packet Switching (OPS).
[7] G. Epps, D. Tsiang, and T. Boures, "System Power Challenges," Proc. Cisco Routing Res. Symposium (San Jose, CA, 2006).
[8] International Technology Roadmap for Semiconductors (ITRS), International Roadmap Committee (IRC), International Technology Working Groups, "International Technology Roadmap for Semiconductors: 2007 Edition," 2007, <http://www.itrs.net/reports.html>.
[9] B. Swanson and G. Gilder, "Estimating the Exaflood: The Impact of Video and Rich Media on the Internet—a 'Zettabyte' by 2015?," Discovery Institute, Jan. 29, 2008.
[10] Telcordia Technologies, "Network Equipment-Building System (NEBS) Requirements: Physical Protection," GR-63-CORE, Mar. 2006.
[11] R. S. Tucker, "Optical Packet-Switched WDM Networks: A Cost and Energy Perspective," Proc. Optical Fiber Commun./National Fiber Optic Engineers Conf. (OFC/NFOEC '08) (San Diego, CA, 2008), paper OMG1.
[12] U. S. Department of Energy, Office of Integrated Analysis and Forecasting, Energy Information Administration (EIA), International Energy Outlook 2008, Sept. 2008.
[13] Verizon, "Verizon First to Set Up Energy Efficiency Standards for Network, Data Center and Customer Equipment," Press Release, June 5, 2008.
[14] M. Walker, Surfing the Green Wave in Telecom, Ovum, May 9, 2008.

(Manuscript approved August 2009)

OLIVER TAMM is the technology director in the Chief Technical Office of the Optics Division and is located in Nuremberg, Germany. His responsibility spans technology assessment and innovation creation for OTN, SDH/SONET, packet, and WDM assets across Alcatel-Lucent's optical portfolio. Prior to that, he held the position of technical manager in the Systems Engineering and Architecture department and was lead architect for LambdaUnite® and universal packet mux (UPM) cross connect systems. He holds an M.S. degree in electrical engineering from the Technical University of Darmstadt, Germany, and multiple patents for packet and optical transmission systems.

CHRISTIAN HERMSMEYER is a distinguished member of technical staff in Alcatel-Lucent's Optics Division's Chief Technology Office (CTO) department and is located in Nuremberg, Germany. He received his M.S. degree in electrical engineering from the University of Dortmund, Germany. Mr. Hermsmeyer began his career in hardware development at Philips Kommunikations Industrie AG in Nuremberg, Germany, and Glasgow, Scotland. His work at Alcatel-Lucent has spanned the areas of ASIC design, system engineering, and transmission architecture definition for packet- and transport integrating systems. Within CTO he led an initiative on 100 Gb/s packet processing.
Mr. Hermsmeyer has authored/coauthored various technical conference papers and holds patents for transmission and data networks. He is a member of the Alcatel-Lucent Technical Academy.

ALLEN M. RUSH is a consulting member of technical staff in Alcatel-Lucent's Optics Division's CTO department and is located in Murray Hill, New Jersey. He received a bachelor of science in electrical engineering from the University of Maryland, College Park, and a master of science in electrical engineering from the University of California at Berkeley. As a past member of the Fixed Access Business Group, he has worked on broadband access systems providing digital subscriber line (DSL), fiber-to-the-home (FTTH), and voice-over-IP. Currently, as a member of the Optics CTO Technology Group, his areas of interest include silicon-based photonics, WDM technologies, and power efficiency in transport networks. ◆