♦ Energy-Efficient Transport for the Future Internet

*Gert J. Eilenberger, Stephan Bunse, Lars Dembeck,
Ulrich Gebhard, Frank Ilchmann, Wolfram Lautenschlaeger,
and Jens Milbrandt*

The emergence of new interactive and peer-to-peer broadband services is fostering the growth of subscriber access bandwidth as well as broadband penetration, resulting in a continuous increase in traffic in metro and core networks by a factor of 10 every five years. State-of-the-art Internet Protocol (IP) based core network architectures are expected to suffer from severe scalability problems with respect to complexity, power, and cost. Novel architectural approaches will be required as a basis for the future converged packet transport infrastructure offering petabit networking capabilities at much lower power and cost. We describe a scalable, future-proof architecture which reduces complexity as far as possible by shifting packet processing to the edges of the network, aggregating traffic into large containers, and applying simple circuit switching whenever possible, preferably in the photonic layer. Novel approaches for optimized traffic management contribute to the simplification of processing, protocols, network control, and management. The expected savings, together with service-driven quality of service (QoS) provisioning, can open new ways for implementing high leverage transport networks and deriving new revenues. © 2010 Alcatel-Lucent.

Introduction: Problem Description

Like the telephone connection in the past, access to the Internet is seen today as a commodity in peoples' lives enabling them to stay in touch with friends or to have access to learning opportunities. Therefore, many countries have issued a political memorandum to enable "Broadband for All."

This is accompanied by a change in user behavior. In the past, users were just passive consumers, whereas now they have turned into active participants motivated by new offers like Facebook*, YouTube*, and Twitter* typically summarized under the headline Web 2.0. This is a continuous process and new trends like the Internet of Things, Ambient Assisted Living, and Smart Senior are already emerging on the coming Web 3.0 horizon.

As a consequence, all parts of the transport network (access, metro, backbone) have to support these applications, implying that much more bandwidth and new control schemes will be needed.

The typical access bandwidth today is in the range of 384 kb/s to 2 Mb/s [10], which will need to be increased to 50 to 100 Mb/s to support broadband applications like

Bell Labs Technical Journal 15(2), 147–168 (2010) © 2010 Alcatel-Lucent. Published by Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/bltj.20446

Alcatel·Lucent

Panel 1. Abbreviations, Acronyms, and Terms

BoD—Bandwidth on demand
CAPEX—Capital expenditure
CBDM—Central bypass decision maker
DeCIX—German commercial Internet exchange
DRAM—Dynamic random access memory
ECN—Explicit congestion notification
FAU—Frame aggregation unit
FS-LC—Frame switch line card
HD—High definition
ICT—Information and communication technologies
IETF—Internet Engineering Task Force
IP—Internet Protocol
IPTV—Internet Protocol television
ITU—International Telecommunication Union
ITU-T—ITU Telecommunication Standardization Sector
MAC—Media Access Control
MIP—Mixed integer programming
MPLS—Multiprotocol Label Switching
MPLS-TP—MPLS-Transport Profile
ODU—Optical data unit
OPEX—Operational expenditure
OTH—Optical transport hierarchy
OTN—Optical transport network
P2P—Peer-to-peer
PCE—Path computation element
PCECP—PCE Communication Protocol
QoS—Quality of service
RAM—Random access memory
RED—Random early detection
SAN—Storage area network
SLA—Service level agreement
TCP—Transmission Control Protocol
VoIP—Voice over IP
VPN—Virtual private network
WDM—Wavelength division multiplexing

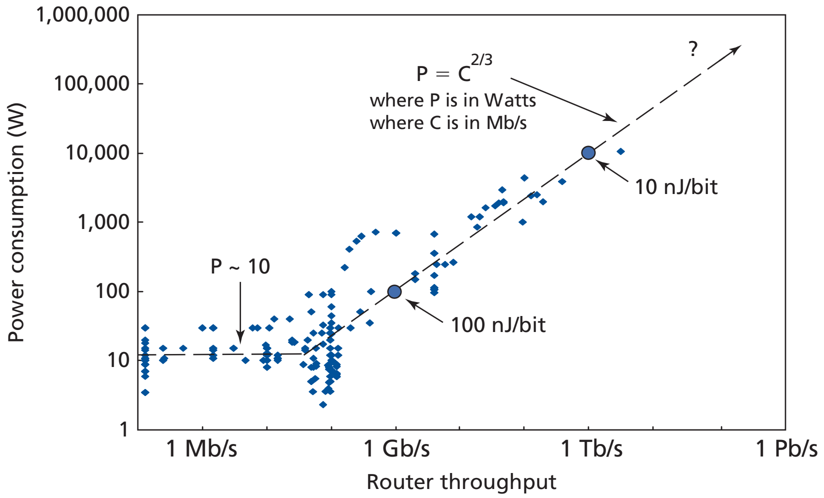
peer-to-peer networking, high definition (HD) video applications, or Internet Protocol television (IPTV). This represents a strong and growing market for many operators. In Germany, for example, only 0.14 percent of the households are connected to IPTV [11].

Even higher access rates in the range of 1 Gb/s are required when considering business services like storage area networks (SANs) and telesurgery. Based on many measurements and forecasts, it is assumed that traffic volumes will grow on average by a factor of 10 within five years, which yields a factor of 100 by 2020 [4, 6, 9, 11]. One example of above-average traffic growth is found in the average and peak traffic through DeCIX, the German Internet exchange. In the ~800 days from July 2007 to July 2009, traffic increased by a factor of four [6].

Current packet networks suffer from quality and availability issues due to heavy traffic contention: if all consumers demanded high bandwidth simultaneously, they would receive only a small fraction of the maximum rate they had booked for their access link. The networks use over-subscription and they rely strongly on statistical multiplexing gains. The advent of new, bandwidth-hungry services will place even greater demands on future transport networks. Video-based services will inevitably require a long holding time, which implies that high statistical multiplexing gains will not be available and contention will need to be reduced to maintain acceptable service quality.

A simple calculation will show the trends: Let us consider a country with 20 million residential customers and 1,000 access nodes, such as Great Britain or Germany; the average number of customers per node is 20,000. If each customer demands 1 Gb/s, assuming a 10:1 concentration factor, the average net demand per access node would be about 2 Tb/s. In networks such as this, 10 to 15 access nodes are connected to one core node resulting on average in add/drop traffic of 20 to 30 Tb/s per core node. Many traffic studies we have performed on deployed and operating core networks have shown that 80 percent to 90 percent of the overall traffic per node is transit traffic. With add/drop traffic representing only 20 percent of the overall traffic in a core node, a total node capacity of 100 to 150 Tb/s will be required.

Future packet transport will require novel network and node architectures to handle these traffic volumes along with techniques to support this high



Source data: METI, 2006, Nordman, 2008

Reprinted with permission from R. Tucker, J. Baliga, R. Ayre, K. Hinton, and W. Sorin, ARC Special Research Centre for Ultra-Broadband Information Networks, University of Melbourne, "Energy Consumption in IP Networks," ECOC 2008.

Figure 1.
Power consumption in routers.

bandwidth efficiently, with respect to both cost and energy. In 2007, the overall power consumption attributed to information and communication technologies (ICT) was judged to be responsible for ~2 percent of global greenhouse gas emissions and to constitute 7 percent of the global energy consumption. Fourteen percent of the ICT energy is attributed to telecommunication network infrastructure [5, 19].

Even if these figures do not appear to be extremely critical at first glance, the power consumption of routers is becoming a bottleneck with the growing traffic, since power consumption of routers is still increasing with throughput, as illustrated in **Figure 1**, despite all of the improvements in semiconductor technologies. In Japan, for example, it is expected that by 2015 routers will consume 9 percent of the nation's 2005 total electricity generation. By 2020 this factor will reach ~50 percent and thus become unsustainable [12, 17].

Electronic routers are facing further processing scalability problems, endangering the future feasibility of packet traffic handling. A simple calculation will illustrate the problem. By nature, IP routers process every single packet in order to route and forward it to its corresponding destination. When using 100 Gb/s links and assuming the shortest possible packet length of 40 bytes, e.g., as for a Transmission Control Protocol (TCP) acknowledgement packet, the router has to process each packet within 5 ns. Increasing the bit rate to 1 Tb/s, the packet processing time will decrease to 0.5 ns. But even if we suppose that Moore's Law will still be applicable to processor speed to allow such fast packet processing, an even more severe problem is the memory bandwidth that is required by the traffic manager for the store-and-forward operations of all packets. Generally, a speed-up factor of 2.5 to 3 compared to the input line rate is required for memory bandwidth, which is composed of a factor of 2 for read/write operations, plus some control overhead, plus some reserve to cope with the speed limitations of the technology used. As an example: a 40 Gb/s line card requires a memory bandwidth of approximately 100 Gb/s. Increasing the line rate to 1 Tb/s will push the required memory bandwidth into the range of 3 Tb/s. According to Samsung [22], a growth rate by a factor of 3.7 over five years in throughput/speed of dynamic random access memory (DRAMs) was observed in the past, which is clearly behind the overall traffic growth rate by a factor of 10 in the same time period.

These figures show that a paradigm shift is required for network architectures and switching concepts in order to sustain the growing traffic rates while limiting and even decreasing power consumption.

Clearly, new approaches are needed to help operators both to define new business models and to reduce capital and operational expenditures by means of drastically simplified network architectures and their operation.

We will describe potential architectural and technological solutions to overcome this dilemma and also indicate possibilities for new business models.

Key Architectural and Technological Approaches

The dilemma facing future broadband transport comprises scalability, complexity, cost, energy, and revenue issues. Thus, there is no unique solution to optimize all areas, but holistic approaches and solutions can be discussed for each aspect of the problem.

In this paper, we focus on transport networking, leaving pure transmission/physical layer optimization issues as well as service layer aspects aside. We have made an attempt to summarize the various approaches under investigation in the research community by formulating three main strategies:

- *Strategy A, "less processing per transported bit."* The history of information transport has shown the trend from circuit-switched, analog plain old telephony service requiring very little energy per subscriber line to digital, multi-Gigabit packet switching with the processing necessary for each individual bit consuming a significant amount of energy. Yes, there is much more information transported today and we can enjoy much better service features and quality per energy unit. Even so, the consumption of energy for packet processing and routing is rising in absolute terms over time and we cannot afford to spend the majority of electrical energy for this alone [18]. As a consequence, we have to find ways to transport packet-based services with much less processing effort and energy.
- *Strategy B, "less idle bits per payload bit."* The energy efficiency of transport networks is directly linked to the number of installed resources, as well as their effective utilization. There are two aspects, the overhead required for transport and routing/switching protocols, as well as the spare resources which have to be provided at the network level to guarantee the required degree of resiliency and quality of service for packet based services. As a consequence, we need concepts for network architectures and network operation requiring a minimum amount of resources (in terms of cost and energy) per transported payload bit.
- *Strategy C, "less energy per processing step."* Energy consumption is increasing with speed and the amount of bits to be processed per time interval. In view of the ubiquitous digital IP-based services, it is obvious that we cannot do without a certain amount of packet processing. As a consequence, we have to focus on all possibilities to leverage semiconductor technologies for reducing feature sizes, increasing efficiency, and minimizing power consumption. Intelligent power and cooling management strategies deactivating unneeded equipment will contribute to this as well.

Building a Concept for Future Packet Transport

The strategies previously described will eventually lead to several architectural variants, and, hence, a discussion of their pros and cons is needed to define the best compromise.

Following Strategy A, "less processing per transported bit," means reducing the processing complexity by choosing a lower layer to transport the information whenever possible while still providing the required routing or switching functionality.

Routing in layer 3, the IP layer, is seen as the most expensive transport service, since each single packet with its full header needs to be processed.

Applying Multiprotocol Label Switching (MPLS) switching to IP packet streams offers the first level of simplification by attaching an additional label for simpler switching and avoiding the full header inspection and table look-up functions for the individual packets. Carrier-grade protocols such as the upcoming

MPLS-Transport Profile (MPLS-TP) are offering even more efficient end-to-end transport. Common to all layer 2 packet traffic transport, however, is the fact that each packet has to be processed individually.

Further simplification can be achieved by aggregating packets to the same destination and of the same service class into large, fixed-length macro-frames, which are then packet switched through the entire core network using label switching. This transport mechanism directly supports the scalability to very high bit rates and avoids the handling of tiny information chunks. Macro-frame switching allows a reduction in the packet header processing rate by at least two orders of magnitude in core networks.

Circuit switching provides a platform for further significant reduction in packet processing. Electronic solutions outlined by International Telecommunication Union Telecommunication Standardization Sector (ITU-T) standard G.709 [13] offer a variety of sub-lambda granularity steps that can be used to map different packet traffic streams inside and transport them through the network. Still, the remaining electronic processing and the optical/electronic/optical conversions at each node represent a certain level of complexity.

The most energy-efficient and least expensive transport solution clearly is the switching of individual wavelengths without any bit or packet processing. Hence, lambda switching should be applied to transport packet streams whenever possible.

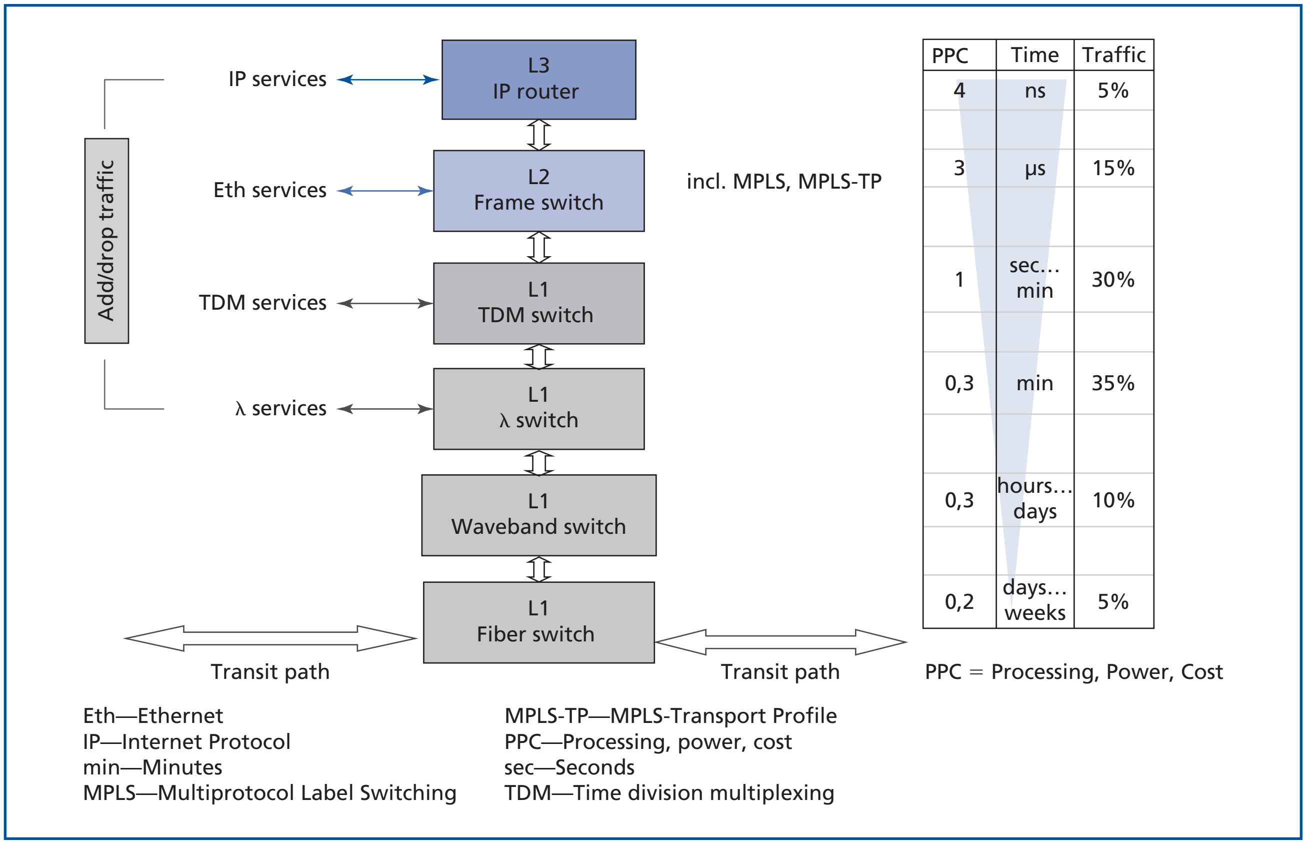
Directing traffic to lower layers, preferably to the optical layer, and thus bypassing complex electronic processing is a possibility for transit traffic, i.e., traffic that is only passing through a node without any need for further processing. The potential for optical bypass is enormous, since 80 to 90 percent of the overall traffic per node is transit traffic.

Figure 2 illustrates the savings potential of different switching and routing technologies. The processing complexity, power consumption, and cost of optical wavelength switching technologies can be more than one order of magnitude lower than the L3 IP solutions. The functionalities offered by the different technologies are decreasing, of course, going from L3 to L1. At the same time, the switching or reconfiguration time will increase for a reasonable exploitation of the related technologies. Therefore, complexity and power savings will not come for free. The technologies must be carefully selected to maintain the overall networking functionalities. Users' end-to-end packet based services must not be compromised, especially when applying the bypassing principle. The estimated traffic share that can be transported at the different layers, however, indicates a high potential for savings by the fact that about 80 percent of traffic could be transported in the circuit switching regime. In the end, the overall network costs and energy needs will decrease dramatically.

A network architecture optimized according to Strategy A would consequently consist of nodes interconnected with a full mesh of wavelength channels, thus switching at the least expensive layer and avoiding any intermediate processing.

Strategy B, "less idle bits per payload bit," favors a different solution. The most efficient way to utilize network resources and to reduce overhead is to exploit the statistical multiplexing as far as possible. This means concentrating as much traffic as possible on a small number of links with huge bandwidth. In order to achieve this, we need to stay mainly in the packet switching layer, which requires a significant amount of packet processing for adding and dropping packets to/from the big bandwidth pipes.

It is obvious that the network architectures resulting from Strategy A and Strategy B are somewhat contradictory. We see two extreme cases, one requiring very simple processing but a huge network of poorly utilized wavelength channels, the other with a small number of optimally used network resources requiring an extremely high processing effort. The optimum will probably be a combination of both strategies, where the packet processing by IP routers is moved to the edge of the network to support packet-based customer services. The core network remains routerless and, in the ideal case, exploits only wavelength switching. Traffic variations and dynamics may request more flexibility in the core, which could be served by sub-lambda switching solutions like optical transport network (OTN) or by ODUflex, a new approach offering fine granularities down to 1.25 Gb/s for flexible bandwidth-on-demand scenarios in the OTN context. All of these electronic sub-lambda



PPC	Time	Traffic
4	ns	5%
3	µs	15%
1	sec… min	30%
0,3	min	35%
0,3	hours… days	10%
0,2	days… weeks	5%

Eth—Ethernet
IP—Internet Protocol
min—Minutes
MPLS—Multiprotocol Label Switching
MPLS-TP—MPLS-Transport Profile
PPC—Processing, power, cost
sec—Seconds
TDM—Time division multiplexing

Figure 2.
Savings potential of different switching/routing technologies.

switching approaches are already being commercially deployed or are in the final stages of standardization; however, they are still circuit switching schemes. A certain fraction of the traffic may still require packet switching capabilities in the core in order to exploit the statistical multiplexing potential for optimized utilization of network resources, or to more easily implement dynamic multicasting schemes. For these cases, at minimum required packet switching functionalities, we are looking in particular at simple and energy-efficient approaches like macro-frame switching.

Another important aspect is support for all service bandwidth granularities that the customers may request across all layers (L1, L2, L3) including express lanes and virtual private networks (VPNs). In an efficient solution, these granularities would be directly supported by tailored network services in a multilayer network.

We have seen that the optimum network architecture will result from a well-balanced composition of many parameters in a multi-layer environment. It is not easily defined in a single attempt but will need several iterations, taking into account technological, service, and business model aspects.

Proposals for Networking Solutions

We will describe some important selected aspects for networking solutions in the sections following.

Cost and Power Optimization of Network Architectures

We will investigate two opposing architecture approaches and thereby consider aspects of eco-sustainability in network optimization.

The first approach is designed according to Strategy A, "less processing per transported bit." The objective is to minimize the transit traffic by setting up a direct light path between every pair of edge devices exchanging any traffic. As a result, no traffic is processed by intermediate

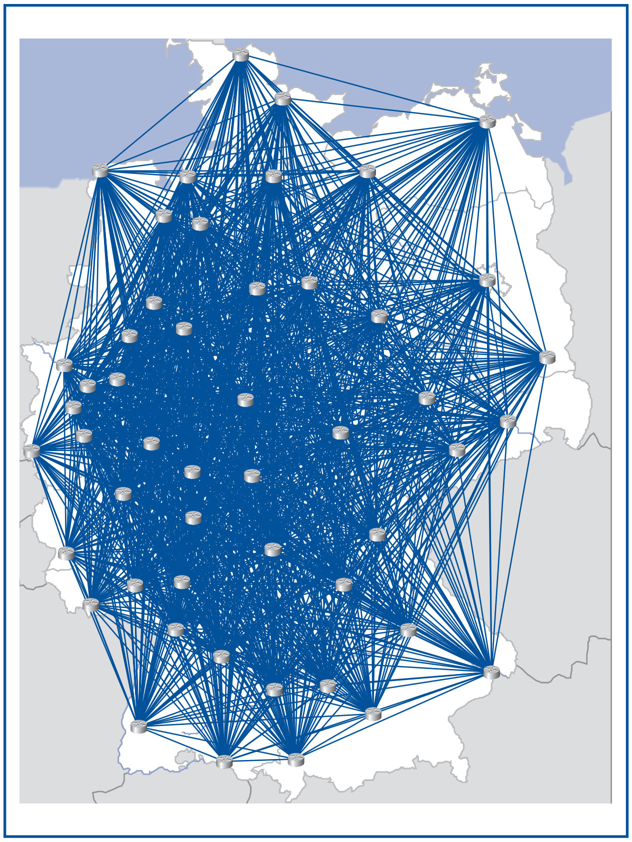


Figure 3.
Strategy A: full mesh logical topology.

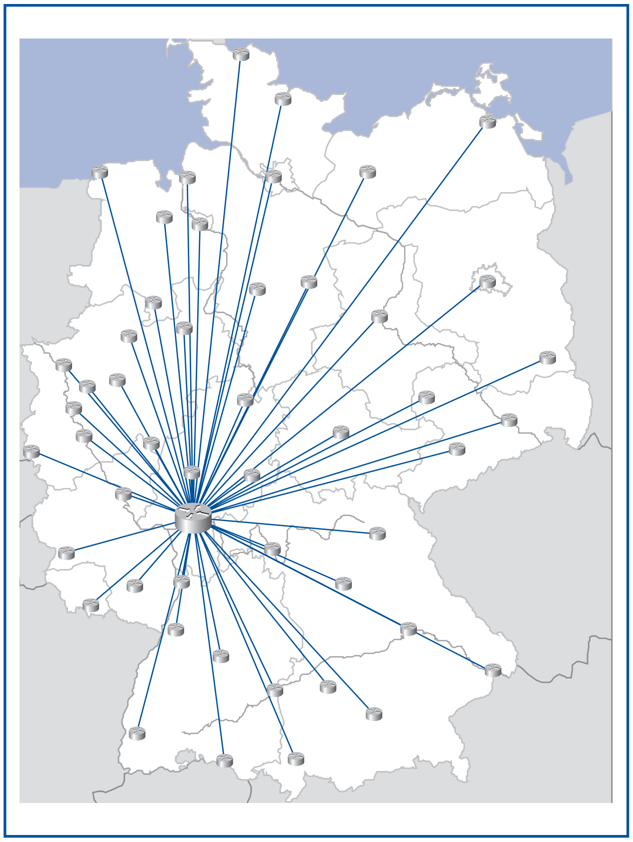


Figure 4.
Strategy B: one star logical topology.

devices, which are instead bypassed in the optical layer. With this, there is also no aggregation of traffic. The approach induces many logical links. The resulting logical topology is a fully connected graph, similar to that shown in **Figure 3**. From the power consumption point of view, no power is required for transit traffic processing, but there will be some power wasted in the many interfaces of the edge devices.

In the second approach (Strategy B), the objective is to have "less idle bits per payload bit." This is done by using as few connections as possible. As a result, all traffic at an edge device is aggregated into one connection directed to a central device, similar to the example in **Figure 4**. This topology will minimize the number of interfaces and their corresponding power consumption at the edge devices but increase the amount of transit traffic in the central device.

Between these two opposite optimization approaches, a tradeoff must be found to achieve an optimum in cost and power consumption. **Figure 5** illustrates the relations between the extreme logical topologies described above. The tradeoff between the two approaches has to be found in the range between the extremes, considering high and low transit capacity as well as low and high port cost.

Optimization methods. Network optimization can be formulated as a mixed integer programming (MIP) problem, which allows for exact results. Alternatively, heuristic procedures can be applied. Network design problems are usually very complex. Therefore, exact methods often need very lengthy computing times and significant amounts of memory. Although computing power as well as algorithm efficiency have increased by orders of magnitude over the past few years, optimal designs can only be provided for rather small networks. This holds especially for the design of multi-layer networks. Heuristic methods are often based on intuitive approaches and can often find a possible node design very rapidly. The intuitive solution is normally not the optimum one, but one sufficiently close to be viable.

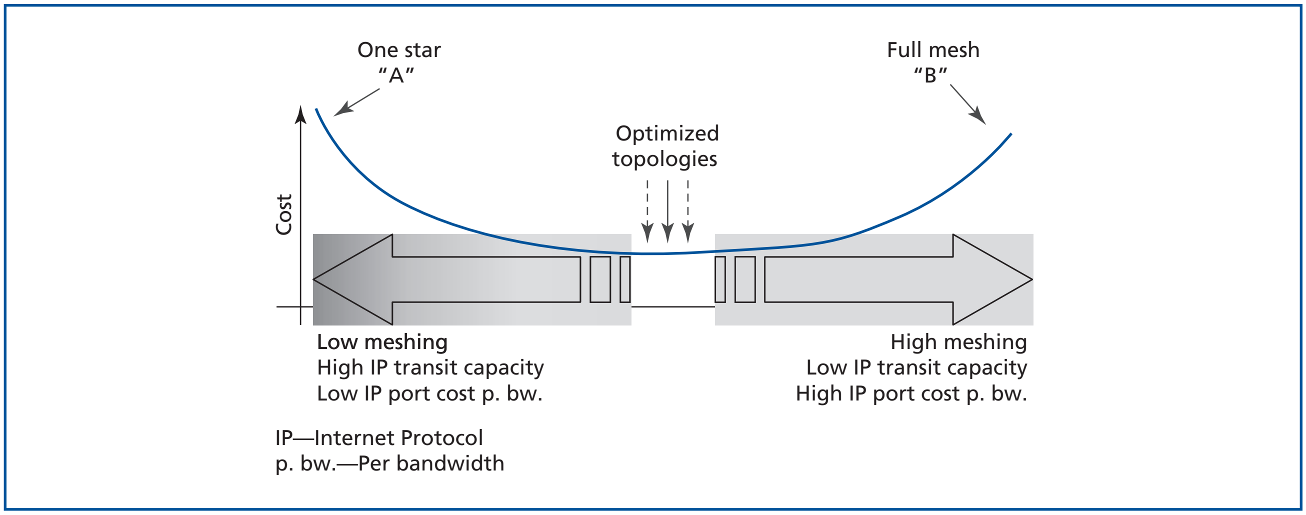


Figure 5.
Relation between meshing, transit capacity and port cost.

A multiplicity of heuristics exist for optimization of logical network topologies. Two algorithm approaches have been validated in detail: composition and decomposition algorithms. These algorithms are well known from literature, deliver satisfying results in an acceptable expenditure of time [23], and have been implemented on a suitable dimensioning platform.

The goal of both algorithms is to reduce the cost of the network by increasing (composition) or decreasing (decomposition) the meshing of the logical network topology, respectively.

An example composition approach can be described as follows: In an initial step, the starting topology is established as a minimum meshed network and a list of candidate links is generated. After the setup of these candidate links, the heuristic runs in a loop, as long as network costs are reduced. The candidate links are included one after the other into the topology. The body of the loop contains the network dimensioning (routing of the demands on the current topology) and the cost calculation (capacity calculation of links and nodes from demand routing). The cost calculation is based on a model that aggregates the cost of each possible device module. **Figure 6** presents a flow diagram of this solution. **Figure 7** shows the result of the composition algorithm for a European backbone network with ~80 nodes in the logical layer and ~140 nodes in the physical layer.

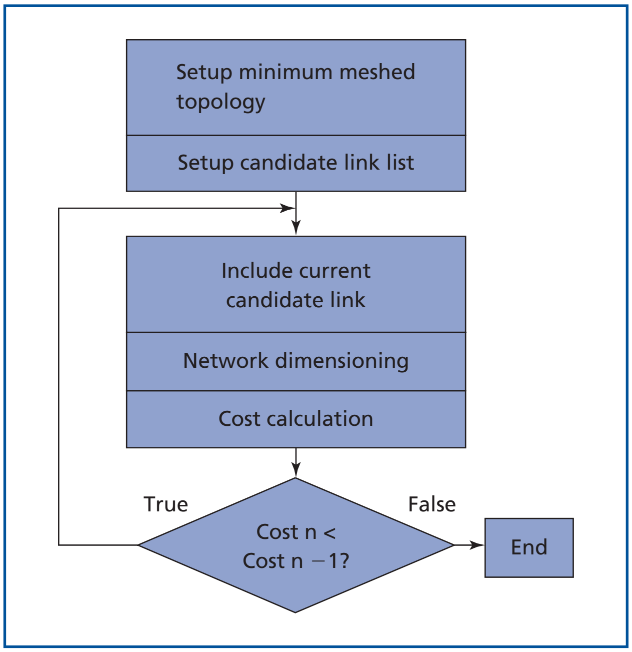

Setup minimum meshed topology
Setup candidate link list
Include current candidate link
Network dimensioning
Cost calculation
True
False
Cost n <
Cost n − 1?
End

Figure 6.
Flow diagram of the composition approach.

In the initial phase of optimization, the cost plunges and afterwards the algorithm runs in a saturation region with minimum cost reductions. The graph also shows that the implementation of the algorithm tolerates a temporary cost increase.

We extend the cost optimization beyond conventional capital expenditure (CAPEX) considerations by

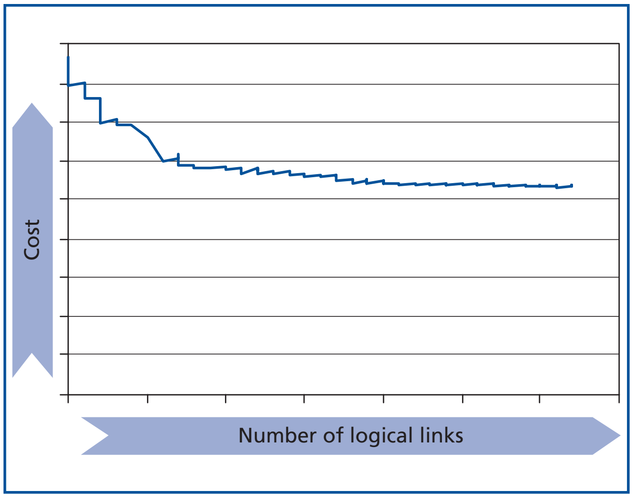


Figure 7.
Overall network cost over number of logical links.

including operational expenditure (OPEX) components:

- Energy consumption,
- Cooling,
- Network operation and maintenance, and
- Location maintenance.

The optimization algorithm uses device models, illustrated in **Figure 8**, to evaluate the impact of device types deployed on network lifetime cost.

The optimization models run up to this point clearly show the potential for savings. Overall cost savings and power reduction in the range of 60 to 80 percent compared to all-electronic layer 3-based solutions (IP routers equipped with clear channel interfaces over a static WDM layer) appear possible by exploiting the potential of photonic circuit switching technologies. However, many further studies have to be performed until all aspects have been sufficiently considered and a final architecture is defined.

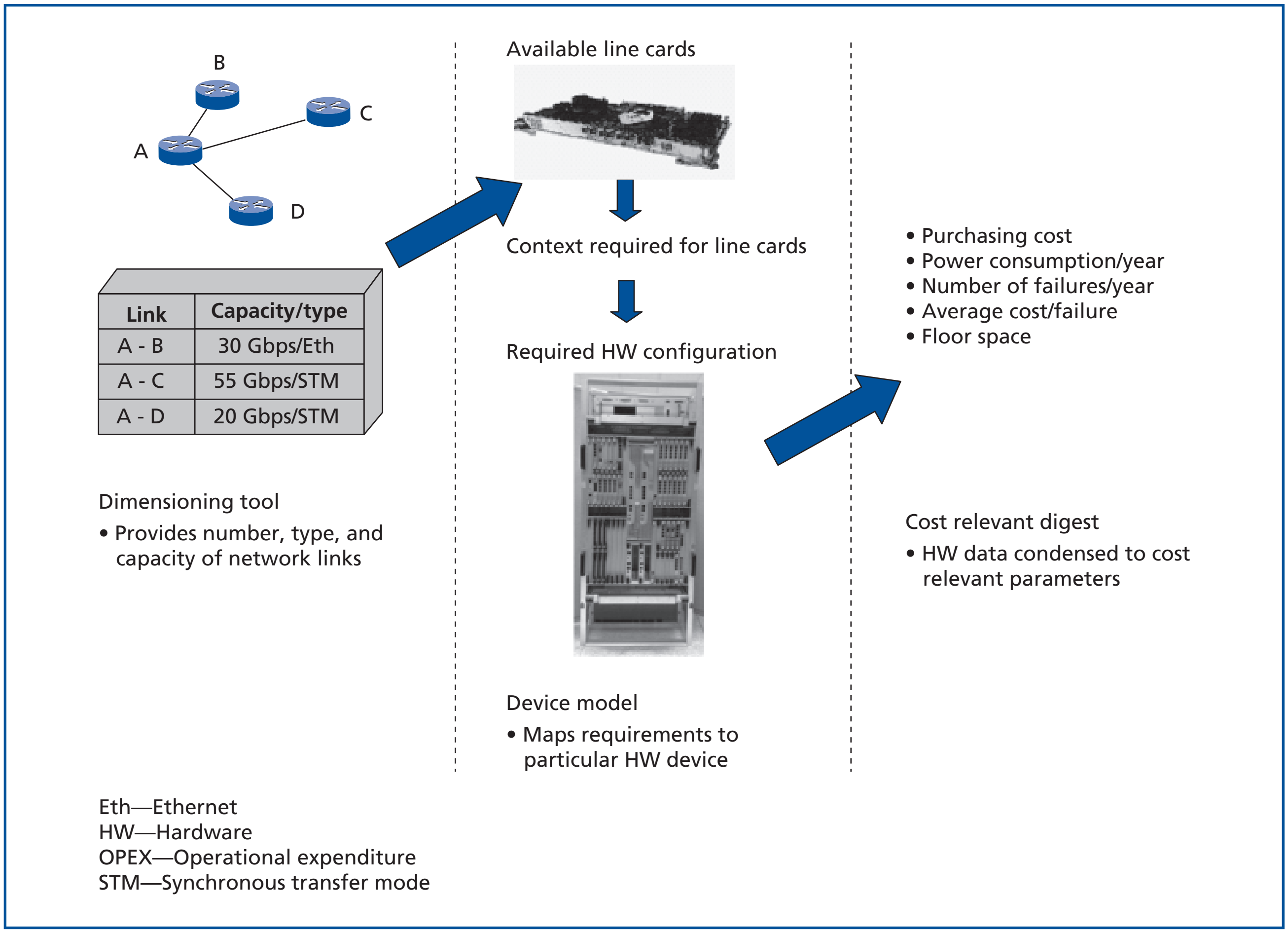
Bypassing Expensive Electronic Processing

In today's core transport networks, most of the IP traffic is transit traffic (in most of the nodes) and, hence, IP packets should be bypassed in sub-IP layers wherever possible to lower the required packet processing power and to save energy.

The principle corresponding to this objective is well known by various names including optical bypassing and light path routing. Whatever the name, the principle remains the same and can be described as follows: We suppose an abstract, hierarchical, multi-layer network with a packet-over-circuit switched architecture using arbitrary packet and circuit switching technologies (e.g., IP over WDM). Packets are transported hop-by-hop through circuits directly connecting each two packet switches along a packet's path: i.e., on its way through the network, each packet is processed in each hop. For a given traffic situation, an optimal layout of circuits is determined with regard to different objective functions and subject to specific technical side conditions. The models and proposed solutions to this optimization problem are numerous [18, 20, 21, 24] and many of them have only pure academic value.

A simple bypass deployment approach. In full awareness of the many previous studies on optical bypassing, we focus on a practical approach that is applicable to today's multilayer networks and comprises all functions necessary for the deployment of bypasses in operational network environments. The basic idea is illustrated in **Figure 9**, where packet switches, e.g., B* and C*, measure and process the transit traffic between neighbor nodes. When that traffic exceeds a pre-defined threshold, e.g., from feasible circuit bypass capacities, the affected packet switch will request a circuit bypass between neighbor nodes exchanging heavy traffic to offload itself from unnecessary packet processing. It therefore pushes a request to a central bypass decision maker (CBDM), which decides on the deployment of a bypass. The CBDM thereby processes concurrent bypass requests originating from any packet switch on the network in a transaction-oriented manner and coordinates the deployment of mutually influencing bypasses (e.g., between previous-/next-hop node pairs A/C and B/D in Figure 9) to prevent any detrimental effects of traffic shifts caused by bypass deployment. In the case of positive acknowledgement, the CBDM arranges a circuit bypass between the neighbor nodes of the requesting packet switch.

The regular forwarding process in a packet switch is analyzed to measure transit traffic between previous-/next-hop node pairs. To do so, packets arriving at a switch are consecutively stored in different buffers on the incoming interface for the purpose of general packet inspection. Based on the destination address



Link	Capacity/type
A - B	30 Gbps/Eth
A - C	55 Gbps/STM
A - D	20 Gbps/STM

B
C
A
D
Available line cards
Context required for line cards
Required HW configuration
• Purchasing cost
• Power consumption/year
• Number of failures/year
• Average cost/failure
• Floor space
Dimensioning tool
• Provides number, type, and capacity of network links
Cost relevant digest
• HW data condensed to cost relevant parameters
Device model
• Maps requirements to particular HW device
Eth—Ethernet
HW—Hardware
OPEX—Operational expenditure
STM—Synchronous transfer mode

Figure 8.
Using device models for OPEX calculation.

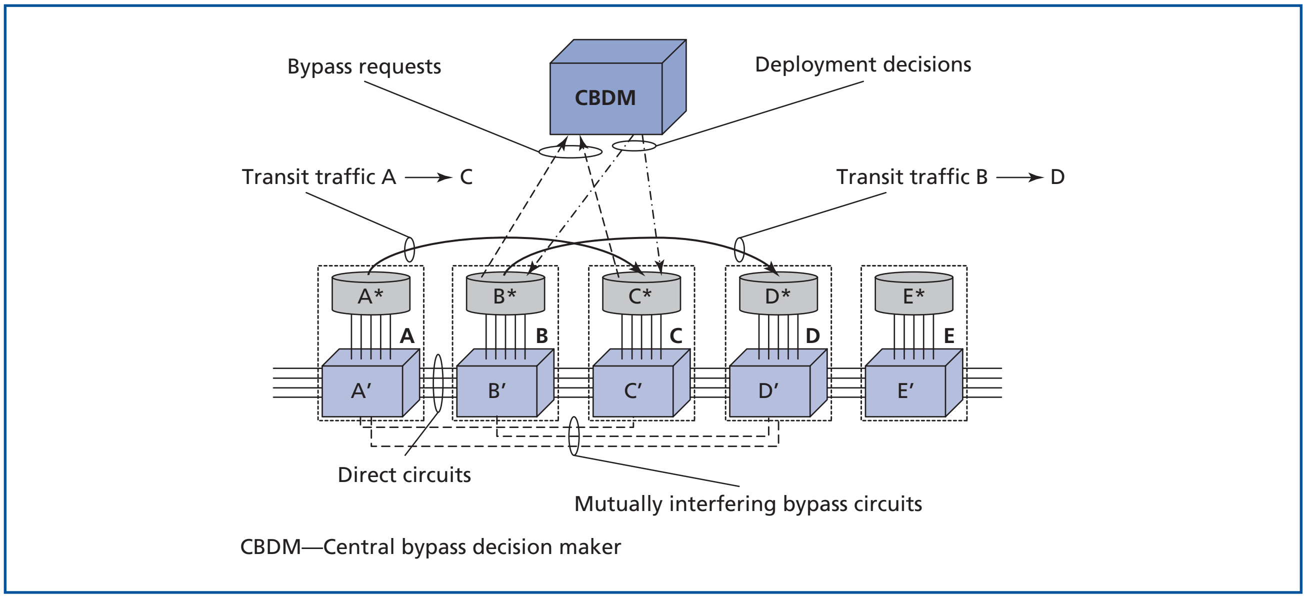


Figure 9.
Rational bypass deployment approach.

of a packet, the usual forwarding table lookup is performed to identify the outgoing interface. At this time, the incoming interface, the packet length in bytes, and the outgoing interface/next-hop node of a packet are known. The previous-hop node of a packet can be identified by looking up the forwarding adjacency table, which is set up by routing protocols and maps interfaces to packet switch neighbors. Accounting for packets per previous-/next-hop node pair and considering their cumulated lengths within a measurement interval enable the calculation of traffic rates. Hence, all necessary transit traffic information, i.e., previous-/next-hop node pair and corresponding traffic rate, for issuing a bypass request is determined.

Intelligent resolution of bypass concurrency. Any packet switch detecting excessive transit traffic can issue a request for bypass deployment to the CBDM managing bypasses, similar to a path computation element (PCE) [7] managing explicit routes in a network. The CBDM has the complete overview of the bypass situation in the network. It resolves the concurrency of simultaneous bypass requests and detects the mutual, possibly detrimental interference between already existing and newly requested bypasses. The contemporaneity of bypass requests thereby is a matter of parameterization of the CBDM. In concurrency situations with simultaneous bypass requests, e.g., issued by B* and C* in Figure 9, either a single bypass, A-C or B-D, is deployed, or both may be deployed. Prior to the enablement of a bypass, the CBDM evaluates its effect on other bypasses as well as on the overall packet processing situation in the network. It thereby resolves interference problems between previously established and newly requested bypasses. If, for instance, bypass B-D in Figure 9 is already deployed and packet switch D* requests a bypass between B* (which becomes a new neighbor of D* after bypass B-D has been deployed) and E*, the impact of potential bypass B-E on B-D is evaluated prior to enablement. If bypass B-E takes over some traffic of B-D, thus rendering that bypass unprofitable, its deployment may or may not be granted depending on its balance of gain and loss of packet processing savings. Similar to the construction of new bypasses, the CBDM may be informed of bypasses that become unprofitable due to traffic changes. It then has to decide whether the teardown of a bypass and the associated reclaiming of resources outweigh the loss of packet processing savings.

A centralized architecture using a CBDM as previously described is only one alternative for a bypass deployment implementation. Its major advantage is an easy decision process based on a global network view. Its implementation, however, requires a new, simple protocol, similar to the Internet Engineering Task Force (IETF) PCE Communication Protocol (PCECP) [1], for signaling between the CBDM and the switching devices. In an alternative distributed architecture, packet switches could use an existing integrated control plane with some necessary extensions for signaling overload to previous-hop nodes, which then decide on the bypass deployment. More signaling is needed afterwards to direct the transit traffic into established bypasses. This, however, requires bringing the bypass deployment intelligence down to the switching devices and raises synchronization issues regarding concurrent bypass requests. For the sake of least complexity, this favors a centralized approach like the CBDM.

Traffic Aggregation to Reduce Packet Processing Complexity

Aggregation of packet traffic into macro-frames and macro-frame switching calls for a specific packet switching technology that uses large transport containers. The frame size could range, e.g., from 9 to 16 Kbytes, which is contrary to the current Internet packet granularity, which ranges from only 40 to 1500 bytes at maximum. Apart from the container size, macro-frame switching is quite similar to packet switching. It uses header information for routing of the individual containers, preferably labels. Containers are transported in a store-and-forward way. Buffers are used in network nodes, where the randomly arriving macro-frames may encounter contention at a common output port. In this way, the ability of packet networks to instantaneously adapt to changing traffic conditions is preserved. The performance of a macro-frame switching network with respect to statistical multiplexing is almost the same as that of packet networks.

The benefits of macro-frame switching. Within packet networks, the price of flexibility, when compared with circuit switching, is the considerable processing effort per packet in every intermediate node along the path to the destination. Whether or not the flexibility pays off is dependent on the actual traffic conditions. At this point, macro-frame switching is an attempt to extend the reach of packet switching by reducing the processing effort but maintaining the same flexibility.

The container frames of 9 to 16 Kbytes are not much larger than the largest Internet packets of today (by a factor of 6 to 10). Nonetheless, the processing power of network nodes can be dramatically reduced. This is due to the fact that ordinary Internet nodes are designed to cope with a worst-case scenario of small packets at heavy loads as well as at light loads. All line card processing functions like classification, address lookup, or memory interfaces must be implemented in a way that arbitrary, long sequences of the smallest packets can be processed on time. There is no averaging with other, possibly larger, packets in the traffic. This does not necessarily mean that the real packet processing rate is that high all the time, but the capability for full processing power has to be there, and it must be available within a few nanoseconds. In the case of macro-frame switching with a smallest frame size of at least 9 Kbytes, however, the required processing power can be relaxed by factor 100 or more.

Implementation. The benefits of macro-frame switching suggest its possible utilization in all networking environments. Packet switching creates an inherent delay at the application level that cumulates with the time that an application needs to fill a single packet before it can be sent out. Voice over IP (VoIP) as a typical narrowband application creates a data rate of 16 to 64 kb/s depending on the codec. It would take 4.5 seconds to fill a 9 Kbyte frame with a single VoIP session, which is unacceptable for voice conversations. In practice, a delay of only 20 milliseconds is permitted. That's why VoIP applications are using packet sizes of only 70 to 250 bytes. Other essential applications using small IP packets include signaling (e.g., TCP acknowledgement packets) and sensor networks. Towards the core networks, the store-and-forward delay for large frames is no longer a problem. At a 10 Gb/s line rate, the transmission of a 9 Kbyte frame takes only 7.2 microseconds, which is negligible since it corresponds to a fiber propagation distance of only 1.4 km.

The contradictory requirements—small packets in the applications, large frames in the network core—can be fulfilled, e.g., by an intermediate stage like a frame aggregation unit (FAU). It receives packet traffic; encapsulates multiple packets into fewer transport containers, the macro-frames; and then forwards the macro-frames to the network core, where they can undergo further acceleration and multiplexing. Frame aggregation has to be applied to the aggregated traffic of many users or applications. Otherwise, if positioned immediately behind a single user or application, it would create the same delay problem as described above.

Furthermore, frame aggregation is not just a simple per-interface mapping of packets into containers for the following reason: Frame aggregation (and its reverse) is typical packet processing at a fine granularity and corresponding effort; it offers no savings in and of itself. Cost and power savings occur if frame aggregation is done at the edges of the core network. Then the macro-frames traverse the network core without intermediate unloading and reloading. Only at the egress edge are the original packets released from their macro-frame containers. To achieve this behavior, packets must be classified beforehand by destination egress node and by service class. In consequence, the FAU assembles the classified packets into homogeneous macro-frames on parallel stages, one per forwarding class.

A more detailed investigation of the macro-frame switching architecture, appropriate dimensioning, performance, and impact on the surrounding packet network together with results of a prototype implementation has been reported in [16].

Technology Aspects for Node Realization

Next, we will discuss in more detail the power savings that can be expected if large, fixed-sized data units (e.g., the macro-frames for the macro-frame switching concept described in the section titled "Traffic Aggregation to Reduce Packet Processing Complexity") are used instead of short, variable length packets like those in IP/MPLS, MPLS-TP, or Ethernet. The short minimum length of single packets places especially high demands on the electronic processing.

In this context, we focused on the attribution of power savings as well, to identify how power savings could potentially be achieved with reduced efforts for standardization.

In the following, we will shortly review the approach taken and describe the building blocks of a frame aggregation unit and a frame switch line card (FS-LC). The differences concerning power consumption and effort for each building block will be described and analyzed. In the end, there will be a comparison for the whole FAU and FS-LC.

Evaluation approach for a comparison of macro-frame switching vs. MPLS-TP. This investigation focuses on the power consumption of the line card (FS-LC) and the frame aggregation unit. Since the overall power consumption of a macro-frame switching network also depends on the kind of routing and aggregation at the network level, the outcome of this analysis is intended as an input for a network-wide simulation.

We will perform the comparison relative to the MPLS-TP protocol since the protocol was designed for carrier grade packet transport much the same as the macro-frame switching approach. The control plane and data plane are separated and there is no need to use layer 3 processing for control packets as with MPLS.

A primary topic in the macro-frame switching approach is imitation of networking functions to only the extent needed, so the number of paths in this approach was intentionally limited to about 5,000 paths. These will be sufficient to handle a network of about 70 nodes with moderate meshing. In general, it is assumed that macro-frame switching offers intermediate processing for networks which are too small or carry too little traffic to allow for efficient meshing at the WDM level, but too big for pure Ethernet internetworking in order to be energy optimal. For this reason, setting the number of paths at 5,000 is considered to be reasonable and sufficient. Going beyond this number would increase the data handling significantly.

The level of abstraction at which the comparison should be done poses an important question. We decided to use the functional block level as appropriate. An investigation at lower levels such as the logical gate level may gain higher acceptance. For example, the power that a gate consumes is better fixed than that of a parser. But, on the other hand, one has to realize that the design needs to be done at the gate level to perform this kind of comparison. For these reasons, we decided to perform the study at a block level.

We divided the FAU and the FS-LC into blocks similar to the ones available on a standard packet line card. The power the function needs on an MPLS-TP line card will be normalized to one and the power for the same function for macro-frame switching will be estimated. The next step is determining the relevance of the function for the contribution to the overall power consumption. The starting values are the distribution of the power dissipation on the MPLS-TP line card; the energy consumption for the additional macro-frame switching functions is added on top.

Structure of the FAU and the FS-LC. In the following, we describe the structure of the FAU and the FS-LC in order to perform a power attribution to the different functional blocks.

Frame aggregation is performed on an add/drop line card known as a frame aggregation unit. It is placed in the frame switch. **Figure 10** illustrates the structure of the frame aggregation unit. The FAU receives Ethernet packets from a router, aggregates them in a macro-frame, and transfers them to the internal matrix of the frame switch. The macro-frames can then be sent to any required output line card.

The Ethernet packets received from the router form a pseudowire carrying MPLS-TP packets. The macro-frame switched paths are set up at the MPLS-TP routers, so here we only have a limited number of paths. Thereafter, editing as well as metering and marking are performed. For frame assembly, the packets are written in a multi-queue buffer. If a macro-frame is filled or a timer is expired and the macro-frame must be released, the buffer is emptied and encapsulated in the macro-frame with the Ethernet Media Access Control (MAC) protocol. Thereafter, the macro-frames must be buffered again before they can be scheduled according to quality of service (QoS) principles through the matrix.

On the reverse side, the macro-frames are terminated and the MPLS-TP packet payload is extracted.

The important point here is that the FAU needs two buffers, one for frame assembly and the other one

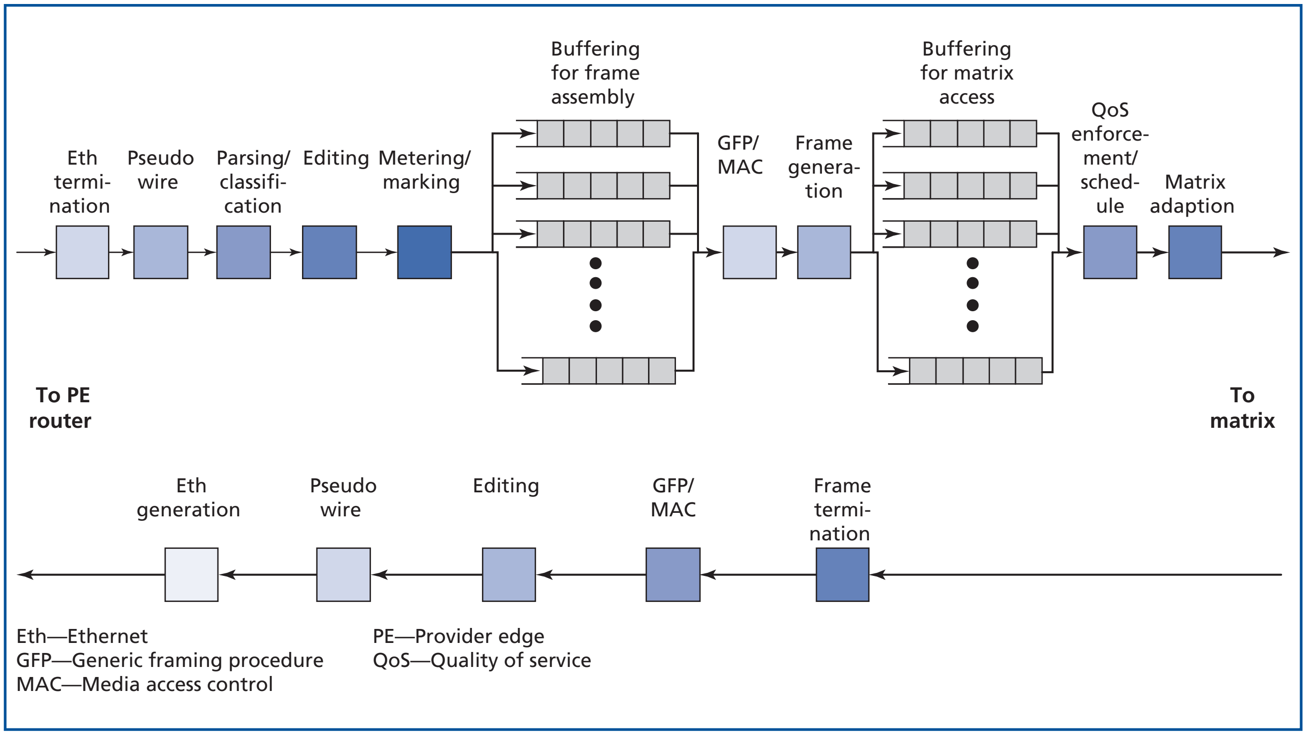


Figure 10.
Structure of the frame aggregation unit.

for the matrix access. As neither can be implemented on-chip, this introduces significant limitations in the FAU design and leads to high power consumption.

The FS-LC provides functions similar to a standard MPLS line card. First, the optical transport hierarchy (OTH) transport signal needs to be terminated. A frame detection function provides the start and end points of the macro-frame. Parsing, classification, and editing follow thereafter. The buffer is subdivided into different queues for each flow. The QoS is enforced by the scheduler, guaranteeing a minimum spacing of the guaranteed traffic. In the backward direction, the macro-frames are compiled and written into an OTH frame.

Functional blocks of the FAU and the FS-LC. In the following we will review all processing functions for macro-frame switching and describe the differences concerning energy consumption for macro-frame switching and MPLS-TP. These are parsing/classification, editing, buffering, and QoS enforcement.

Parsing means the extraction of bits from the incoming data unit to form a key, which is used during the classification procedure to assign the packet or macro-frame to a flow. Parsing and classification can become quite expensive if there are many hierarchies or even if stateful classification is applied. Macro-frame switching relies on a single hierarchy and a small number of paths, which allow performing these functions with on-chip memory.

As there is only a single hierarchy to administer, the editing function for the macro-frame switching protocol simply involves a label swapping. In subsequent blocks, all packets or macro-frames of a flow are handled in the same manner. With the low number of labels in mind (5,000 are assumed as a sufficient number) the labels could be either stored on a frame processing chip or could share an external static random access memory (RAM) with other frame processing functions.

Buffering macro-frames is easier than queuing variable length packets like MPLS-TP. It may not be obvious at first glance, but handling large fixed frames allows some optimization. First, no processing is required to split the macro-frames into a variable number of fixed sized pieces for storage. The macro-frames already have a fixed size and their length makes them ideal for saving in burst-oriented DRAMs. Ethernet or IP/MPLS packets are transformed into fixed-sized cells (e.g., 128 or 256 bytes) adding additional overhead, which needs to be transferred and stored in the DRAMs, leading to a higher required memory bandwidth. While this may not be very significant, the large size of the macro-frames results in fewer addressable memory positions. Overall administration becomes easier because of the larger size of the memory frames, with fewer frames negating the need for external storage for an empty slot list.

Macro-frame switching supports two different service classes, "best effort," which is always second priority, and "guaranteed traffic." The latter has to have a minimum spacing between its frames, so that the traffic does not block other guaranteed traffic flows.

Although it has been shown that this only has to be assured on the FAU cards, doing so would mean a high risk—a single misconfiguration or failure in an FAU would mean that thousands of flows could be affected. Therefore, it is advisable to check the guaranteed traffic streams at each intermediate node. The QoS mechanism for macro-frame switching is significantly easier than that for MPLS-TP since only one memory access is required. A two-color marker requires two memory accesses. Further, the low number of paths to be supervised reduces the effort significantly.

Power comparison. Compiling all the points from the last section, we end up with the result depicted in **Figure 11**. The figure shows the power dissipation comparison between an FS-LC and a line card for MPLS-TP on a bit/s basis. Since the application areas of macro-frame switching are core networks, it is assumed that the filling ratio of the frames is close to 1.

The power needed for processing the transport signal remains the same for both line cards; the same is true for the matrix adoption and the on-board controller. Editing, metering, and parsing/classification require significantly less power due to the longer frames used in macro-frame switching. Here the significantly lower number of memory accesses comes

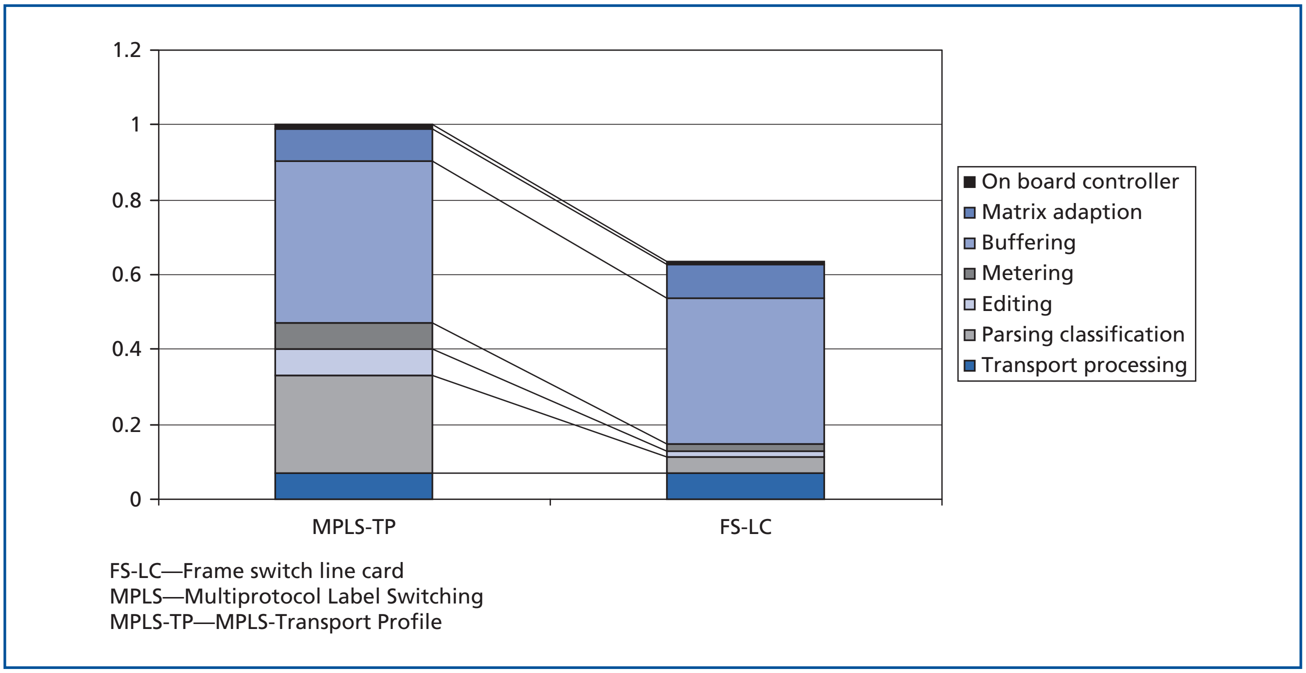


Figure 11.
Power dissipation comparison between a FS-LC and a line card for MPLS-TP.

into play. On the other hand, the power reduction associated with buffering is less significant. About the same memory bandwidth is needed for macro-frame switching and MPLS-TP—only the buffer administration becomes easier. Thus, only a moderate power reduction of 20 percent can be expected. The total calculated power for a line card with macro-frame switching is at 63 percent of the power needed to operate an MPLS-TP line card.

The same investigations were applied to an FAU card, which showed power consumption equal to an MPLS-TP line card. The main point here is that an additional data buffer is needed for compiling the MPLS-TP packets into a macro-frame, and this is responsible for the additional power draw compared to the FS-LC.

Conclusions on the macro-frame switching assessment. We have shown with the power analysis that macro-frame switching is an interesting option for networks which cannot fully rely on circuit-switched optical or electronic OTH-type interconnections but require the flexibility that only packet switching can provide. Unfortunately, the power demand of the line speed buffers does not decrease, which hinders the prospect of a further reduction in energy use. A frame switch line card will consume about one-third less power than an MPLS-TP line card, while the FAUs at the end of the macro-frame switched path require the same energy as MPLS-TP line cards. This makes it advisable to set up frame switching paths, even short ones of one hop or more. Macro-frame switching is very efficient compared to other packet switching technologies like IP, which is estimated to require four times more energy than OTH-type circuit switching schemes. Macro-frame switching is also considerably more efficient than MPLS-TP, which can benefit from a reduced number of table look-ups compared to IP. Nevertheless, we will not reach the OTH energy level even with this efficient technology.

Simple End-to-End QoS Provisioning

Cross layer optimization, bypass deployment, and frame aggregation are mainly directed to Strategy A, "less processing per transported bit," but are somehow contradictory to Strategy B, "less idle bits per transported bit," so there is a need for a tradeoff between the two. In the following, we demonstrate an example with more emphasis on Strategy B, where the focus is not on the carried traffic, but on the unused resource overhead.

Bulk QoS. A major cause of wasted capacity is the permanent request for various kinds of reservation and prioritizations. There are honest reasons to reserve capacity for some particular users and/or applications. Nevertheless, reservations establish artificial boundaries within the resource pool that cause additional opportunities for violations. In general, traffic fragmentation caused by reservations is bad for the efficiency of statistical multiplexing. One way to better utilize resources is to relax the need for reservations and prioritizations. Commonly summarized under the term "bulk QoS," these methods establish a certain level of quality for aggregated traffic as a whole with no regard to the particular traffic fractions.

Why reservations are bad for the statistical multiplexing gain. Reservation is commonly seen as a quality tool. In its broadest sense, it gives sensitive traffic a certain privilege over the rest of the traffic flow. This way, in case of a resource deficit, the sensitive traffic is protected, while the unavoidable losses are assigned to the other traffic. This concept has been explored in numerous ways, with both exclusive and non-exclusive resource reservation, and with hard prioritization or weighted (gradual) privileges. Even the installation of exclusive channels, wavelengths, or links for particular traffic can be seen as a kind of reservation. There are many successful applications of reservation, for example, the prioritization of VoIP over Web traffic on customer access links. Leased lines in an operator network for private network interconnects are another example of exclusive resource reservation. Nevertheless, capacity reservation should not be viewed as a general tool for network quality, and, if applied, it must be understood that one has to pay for it with poor resource utilization overall. The reasons are the following:

- *Prioritization will only work if the privileged class is small compared to the remainder*. If its share becomes dominant in a given traffic mix, then congestion will occur inside the privileged class. The small remainder will be so heavily affected by losses that the "unprivileged" service class will become unviable.

- *Congestion inside reservation bandwidth channels*. Statistical multiplexing relies heavily on the law of large numbers. In terms of network traffic it states: The more components (flows, streams, connections, or packets) are contributing to the actual traffic load, the better the convergence in the long-term mean load. In other words, random fluctuation beyond the mean load will be inconsequential if the number of contributors is large. Reservation will partition the traffic into fractions, each containing fewer contributions than the whole. Convergence to the mean is then compromised, and the overall congestion probability is higher. In the end, considerably more spare bandwidth reserves are needed to cope with fluctuations than would be needed with no reservation.
- *Call blocking of reservations*. The bandwidth on demand (BoD) model offers the potential for congestion mitigation within reserved bandwidth channels. It assumes a reservation per single application session, by which inner congestion is excluded by definition. The problem of BoD services is the outer congestion of call attempts among each other. It has been shown in [14] that the call blocking probability of a BoD service is approximately the same as the packet loss probability presented by a comparable unreserved service, assuming the same load, capacity, and service granularity. A BoD service requires remarkable fluctuation reserves, most likely more than what would be required without reservation.

As a result, resource reservation and prioritization will not improve overall network utilization. In contrast, due to the traffic fragmentation, the statistical multiplexing gain will decrease. A highly fragmented network will need much more fluctuation reserve than a comparable network without reservations.

Nevertheless, there are two reasons to implement reservations:

1. As seen from technical perspective, there are situations where greedy and robust services like Web browsing or peer-to-peer (P2P) services simply squeeze-out more sensitive services like VoIP. Without prioritization, the sensitive service could not be implemented at all.
2. As seen from an economic perspective, network operators want to provide better quality to customers, who are willing to pay more for better service quality. Reservation and prioritization are the only business models currently available.

Coexistence of services on shared resources. Both reasons for reservations can be relaxed by a kind of "bulk QoS." The idea is simple: Aggregated traffic that fluctuates well below the capacity limit does not exhibit considerable losses. Buffers in network nodes are there to absorb contention between simultaneously arriving packets on different interfaces, but no more. Queues, if not empty, are short, and the corresponding packet delay is limited. In such circumstances, the mixture of services doesn't matter. Everything is served on time, up to a remaining statistical uncertainty. Without the need for traffic fragmentation, the overall fluctuation overhead can be better contained than it would be with one of the various reservation schemes.

The crucial question in the idea now on the drawing board is an appropriate measure for "well below the capacity limit." This measure depends not only on traffic load but also on traffic volatility. Various attempts have been made to quantify this measure by the degree of congestion encountered by current traffic. But this includes a problem in principle: Once congestion occurs, it is too late—the quality is already poor. In the opposite case, if no congestion is present, most of the measures are undefined. As an example, we look to the random early detection (RED) algorithm [8]. Its basic premise is to slow the sending TCP end points in a case where the mean queue size exceeds a certain threshold. The mean queue filling ratio is used here as a congestion measure. Under normal working conditions, however, the queue is empty almost all the time. Considerable queue filling ratios will build only during periods of peak congestion, i.e., in cases where the queue size is rapidly approaching the buffer limit. As a result, mean buffer filling is indeed a measure of congestion, but in pre-congestion conditions, it is close to zero and offers no insight relative to the total capacity limit.

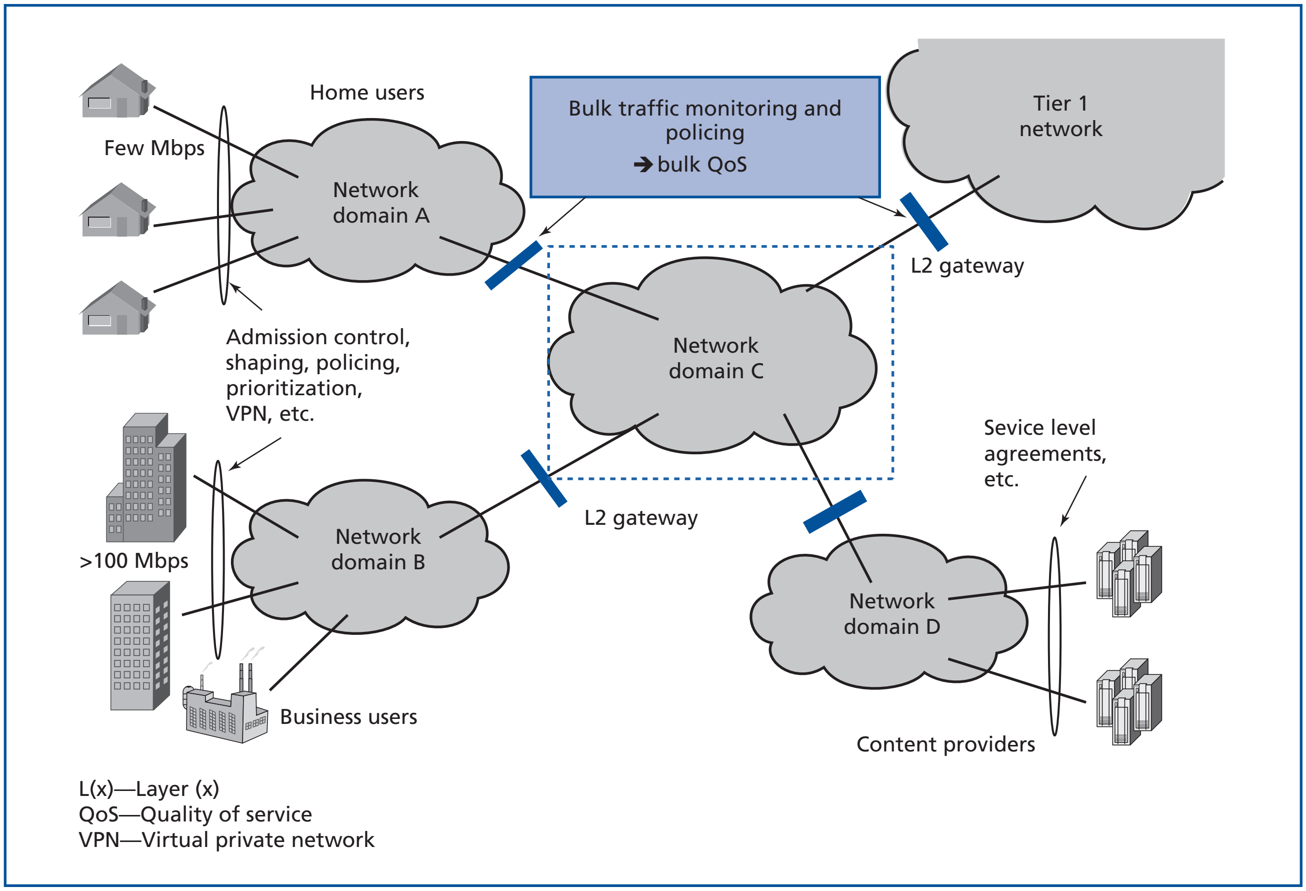


Figure 12.
Principle of end-to-end bulk QoS provisioning.

We follow a different approach for the quantification of congestion. We could show that the volatility of packet traffic is mainly dependent on the bit rate of the contributing application streams [15]. The bit rate of a particular application stream in turn cannot exceed the installed end user access link capacity. This way, a network operator can predict the load and the degree of fluctuation around it by simple observation of the current load together with the knowledge of the installed end user access link capacities. The congestion level can be estimated long before first real losses occur.

Unfortunately, shaping of application streams as a function of access link capacity is not suitable in all circumstances. On the one hand, there are special purpose networks with application streams far below the access link capacity, for example, VoIP on a fast Ethernet infrastructure. On the other hand, there are network domains that do not own the access links but that receive aggregated traffic from other network domains of unknown genesis. In both cases, the granularity of the application stream is not really under operator control. We propose new gateway functions, shown in **Figure 12**, which measure the effective application stream granularity of aggregated traffic crossing a network interconnection point. With this measurement at hand, network operators can manage the traffic volatility in their network domain, agree with other networks on volatility parameters of traffic injections using service level agreements (SLAs), and carry-out corresponding tariffing and policing.

The business model of network domains with managed traffic statistics is still limited to a closed group of network domains with a dedicated purpose and corresponding mutual SLA contracts. It is not applicable to the open character of the Internet. Here, another macroeconomic model comes into consideration [3]. It is built on the assumption that in an open network like the Internet, congestion cannot be prevented beforehand. But, with a kind of congestion pricing, the money flow can be directed from the traffic sources which are creating most of the congestion to the parts of the network which are most congested. With the right incentives, there should emerge an economic equilibrium between demand (traffic offer) and supply (network capacity). Reasonable effort is under way in the context of IETF standardization to implement corresponding mechanisms under the term re-ECN [2].

Conclusions and Outlook

Traffic on the Internet is expected to grow by a factor of 10 every five years. This will raise fundamental challenges regarding network scalability, cost, and energy consumption if the current electronic packet switching paradigm continues to be applied to future packet transport networks as it is today. To overcome these challenges, we have formulated three basic strategies. Besides the ongoing progress in semiconductor technologies, we must achieve "less processing per transported bit" and "less idle bits per payload bit" through the design of novel, scalable, and energy-efficient network architectures. The resulting architecture approaches derived from these objectives are somewhat contradictory and therefore require a quite complex optimization process with numerous parameters to find the best architectural solution. Such an optimized architecture will finally offer a compromise between the extreme cases of a fully meshed, lambda-switched network and a highly concentrated star-type network with a huge central electronic node. It will shift the expensive (and energy intensive) electronic packet routing to the edges of the network and will have a routerless, circuit-switched optical core. Frame aggregation and switching as well as bypassing through electronic sub-lambda (OTH) or optical circuits will be used as much as possible to avoid the costly and energy-hungry electronic packet processing functions. In conjunction with the improvements in semiconductor technology following Moore's Law, we can expect remarkable cost and power savings up to an order of magnitude by applying an optimized architecture today, while the ultimate solution is under investigation.

Among the various parameters and principles, we selected a few key enablers and elaborated on them in this paper. We described a method to implement the bypassing mechanism in a multi-layer network environment. We further described the principle of traffic aggregation into large macro-frames, which reduce packet processing complexity by a factor of more than 100, and we illustrated the possible complexity and power savings in an implementation example. Finally, we presented a new concept for end-to-end QoS provisioning which does not require the monitoring of individual flows and their behavior and thus simplifies the operation of packet networks considerably. The approach includes a simple mechanism for bulk traffic monitoring and policing and offers new business models such as "charge for congestion" and "pay for quality" to operators who are keen to find new sources of revenue.

*Trademarks

Facebook is a trademark of Facebook, Inc.
Twitter is a trademark of Twitter Inc.
YouTube is a registered trademark of Google, Inc.

References

[1] N. Bitar, R. Zhang, and K. Kumaki, "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)," IETF RFC 5376, Nov. 2008, <http://www.ietf.org/rfc/ rfc5376.txt>.

[2] B. Briscoe, A. Jacquet, T. Moncaster, and A. Smith, "Re-ECN: Adding Accountability for Causing Congestion to TCP/IP," IETF Internet Draft, July 14, 2008, <http://www.bobbriscoe.net/ projects/refb/draft-briscoe-tsvwg-re-ecn-tcp-06.html>.

[3] B. Briscoe and S. Rudkin, "Commercial Models for IP Quality of Service Interconnect," BT Tech. J., 23:2 (2005), 171–195.

[4] Cisco Systems, "Cisco Visual Networking Index—Forecast and Methodology, 2007-2012," White Paper, June 16, 2008.

[5] The Climate Group and Global e-Sustainability Initiative(GeSI), SMART 2020: Enabling the Low Carbon Economy in the Information Age, 2008, <http://www.smart2020.org>.

[6] DE-CIX, German Internet Exchange, "Traffic Statistics," <http://www.de-cix.net/content/network/Traffic-Statistics.html>.

[7] A. Farrel, J.-P. Vasseur, and J. Ash, "A Path Computation Element (PCE)-Based Architecture," IETF RFC 4655, Aug. 2006, <http://www.ietf.org/rfc/rfc4655.txt>.

[8] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," IEEE/ACM Trans. Networking, 1:4 (1993), 397–413.

[9] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva, "The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011," IDC White Paper, Mar. 2008.

[10] Germany, Federal Ministry of Economics and Technology, The Federal Government's Broadband Strategy, Feb. 2009.

[11] Germany, Federal Network Agency, Annual Report 2008, Mar. 18, 2009.

[12] H. Imaizumi and H. Morikawa, "Directions Towards Future Green Internet," Proc. 12th Internat. Symposium on Wireless Personal Multimedia Commun. (WPMC '09) (Sendai, Jpn., 2009).

[13] International Telecommunication Union, Telecommunication Standardization Sector, "Interfaces for the Optical Transport Network (OTN)," ITU-T Rec. G.709, v3.4 (draft), Oct. 2009, <http://www.itu.int>.

[14] W. Lautenschlaeger, "Equivalence Conditions of Buffered and Bufferless Network Architectures," Proc. 9. ITG-Fachtagung Photonische Netze (Leipzig, Ger., 2008).

[15] W. Lautenschlaeger and W. Frohberg, "Bandwidth Dimensioning in Packet-Based Aggregation Networks," Proc. 13th Internat. Telecommun. Network Strategy and Planning Symposium (Networks '08) (Budapest, Hun., 2008).

[16] W. Lautenschlaeger, A. Mutter, and S. Gunreben, "Frame Assembly in Packet Core Networks—Overview and Experimental Results," Proc. 10. ITG-Fachtagung Photonische Netze (Leipzig, Ger., 2009).

[17] S. Namiki, T. Hasama, M. Mori, M. Watanabe, and H. Ishikawa, "Dynamic Optical Path Switching for Ultra-Low Energy Consumption and Its Enabling Device Technologies," Proc. Internat. Symposium on Applications and the Internet (SAINT '08) (Turku, Fin., 2008), pp. 393–396.

[18] E. Oki, K. Shiomoto, D. Shimazaki, N. Yamanaka, W. Imajuku, and Y. Takigawa, "Dynamic Multilayer Routing Schemes in GMPLS-Based IP+Optical Networks," IEEE Commun. Mag., 43:1 (2005), 108–114.

[19] M. Pickavet and R. Tucker, "Network Solutions to Reduce the Energy Footprint of ICT," Proc. 34th Eur. Conf. on Optical Commun. (ECOC '08) (Brussels, Bel., 2008), Symposium.

[20] R. Ramaswami and K. N. Sivarajan, "Design of Logical Topologies for Wavelength-Routed Optical Networks," IEEE J. Select. Areas Commun., 14:5 (1996), 840–851.

[21] M. Ruffini, D. O'Mahony, and L. Doyle, "Optical IP Switching for Dynamic Traffic Engineering in Next-Generation Optical Networks," Proc. 11th Internat. Conf. on Optical Network Design and Modeling (ONDM '07) (Athens, Gr., 2007), pp. 309–318.

[22] Samsung, "DDR3 SDRAM," July 2007, <http://www.samsung.com/global/business/semiconductor/products/dram/Products_DDR3SDRAM.html>.

[23] L. Schade, "Technik und Technologie des Fernmeldewesens, 14. Lehrbrief: Optimierung von Telekommunikationsnetzen, Teil Optimale Netzkanten," Zentralstelle für das Hochschulfernstudium, Dresden, Ger., 1991, pp. 48–52.

[24] K. Zhu and B. Mukherjee, "On-Line Approaches for Provisioning Connections of Different Bandwidth Granularities in WDM Mesh Networks," Proc. Optical Fiber Commun. Conf. (OFC '02) (Anaheim, CA, 2002), pp. 549–551.

(Manuscript approved April 2010)

GERT J. EILENBERGER is head of the Packet Transport Networking Technologies Department at Bell Labs in Stuttgart, Germany. His primary research focus is on future high-speed packet transport networks and their control. Dr. Eilenberger's work experience includes synchronous transfer mode (STM) and asynchronous transfer mode (ATM) switching systems; concepts and architectures for optical core and metro transport networks based on WDM and burst/packet techniques; optical and opto-electronic switching and *routing systems; operations, administration, and maintenance (OAM) and control/management concepts; and various system experiments for feasibility verification in the frame of many German national and European research projects. He studied communication engineering and received Dipl.-Ing. and Dr.-Ing. degrees, both from the University of Stuttgart. Dr. Eilenberger has authored more than 55 technical papers on electronic and optical broadband telecommunications and holds 10 patents. He is a member of the Alcatel-Lucent Technical Academy.*

STEPHAN BUNSE is a research engineer in the Packet Transport Networking Technologies Department at Alcatel-Lucent Bell Labs in Stuttgart, Germany. He received his Dipl.-Ing. degree in electrical engineering from the University of Dortmund, Germany. His research interests include communication network architecture and transport network hardware.

LARS DEMBECK is a case engineer at Alcatel-Lucent Bell Labs in Stuttgart, Germany, and he leads a team on optical systems and network activities. He studied electrical engineering with focus on communications engineering and graduated as a Dipl.-Ing. from University of Wuppertal, Germany. Prior to his current posting in Bell Labs, Mr. Dembeck worked for two years in the Alcatel-Lucent software development department. His experience extends from development and implementation of novel electronics for optical network elements in several field trials to the design of concepts and management for photonic networks.

ULRICH GEBHARD is a research engineer in the Packet Transport Networking Technologies Department at Alcatel-Lucent Bell Labs in Stuttgart, Germany. He received the Dipl.-Ing. degree in electrical engineering from the University of Stuttgart, Germany. He began his career at Alcatel-Lucent 19 years ago, focusing first on the development of metropolitan area networks. Five years later, he moved to the optical networks division, where he worked as a senior network engineer for data communications with a focus on the control plane. His experience includes synchronous digital hierarchy/plesiochronous digital hierarchy (SDH/PDH), layer 2 transport technologies, and network protocols. He has been with the Packet Transport Networking Technologies Department for the past six years. His field of activity covers the simulation, dimensioning, and optimization of multi-layer networks.

FRANK ILCHMANN is a hardware development engineer in the Networking and Networks Research Domain at Alcatel-Lucent Bell Labs in Berlin, Germany. He received his Dipl.-Ing. degree in electrical engineering from Technical University Ilmenau, Germany. His field of activity covers the dimensioning and optimization of multi-layer networks with emphasis on data plane simulation tools.

WOLFRAM LAUTENSCHLAEGER is research engineer at Alcatel-Lucent Bell Labs in Stuttgart, Germany, with a research focus on packet transport network and node architectures. His work experience includes theoretical studies on the dynamic behavior of communication systems and their parts, engineering of high-speed signal processing in optical burst mode switching systems, and prototype implementations. He received a Dipl.-Ing. degree in automation engineering from the Saint Petersburg Electrotechnical University (LETI), Russia, and a Dr.-Ing. degree in communication engineering from the University of Transportation and Traffic Sciences "Friedrich List" in Dresden, Germany. Dr. Lautenschlaeger has authored more than 20 technical papers on electronic design and telecommunications and he holds 5 patents together with more than 10 pending filings. He is member of the Alcatel-Lucent Technical Academy.

JENS MILBRANDT is a scientist and network engineer at Alcatel-Lucent Bell Labs Germany in Stuttgart. He received his diploma degree in computer science and economics, as well as a Ph.D. in computer science with a focus on performance evaluation and resource management in communication networks from the University of Wuerzburg, Germany. His research interests are now focused on the evaluation and analysis of new network technologies for packet transport in next generation IP networks. ◆