Traffic-Modeling-Based Design and Evaluation of a Hybrid Electro-Optical Intra-Data Center Network

Mihail Balanici, and Stephan Pachnicke
Institute of Electrical Engineering and Information Technology, Chair of Communications
Christian-Albrechts-Universität zu Kiel, Kaiserstrasse 2, 24143 Kiel, Germany
mihail.balanici@tf.uni-kiel.de, stephan.pachnicke@tf.uni-kiel.de

Abstract
The paper is dealing with the analysis, design and performance evaluation of a hybrid electro-optical Intra-Data Center Network, based on the Long-Range Dependence (LRD) traffic-generation model, as well as the so called “Fat-Tree” architecture. A thorough simulation of the data generation at Server level is performed, followed by the aggregation of data streams at the Top-of-Rack, Aggregation and Core Switch levels. An important contribution in this context is the differentiation of bursty, short traffic patterns from bulky, longer-lasting data streams, which are routed through Electrical Packet Switches (EPS) and Optical Circuit Switches (OCS), respectively. Based on the obtained results, a subsequent CapEx and OpEx cost evaluation is conducted.

1 Introduction
Today’s cloud-based services, such as social networking, e-mail, data storage, and video streaming have reached an unprecedented growth, where a staggering amount of information is generated on a daily basis. That is more than 4 million hours of video content uploaded to YouTube, 4.3 billion Facebook messages and 500 million tweets sent each day. According to recent predictions, by 2020 the Internet traffic will approach 250 exabytes per month [1]. Furthermore, besides the human-user generated data traffic, the Internet-of-Things (IoT) applications are expected to produce a further 3.9 exabytes of information by 2017 resulting from Machine-to-Machine (M2M) communications. All these online activities generate a huge amount of data, which has to be stored, processed and transmitted either within the cloud, or to the client location. Since all cloud applications run and operate in specially designated data centers (DCs), these need to provide a high storage capacity, scalability, and most importantly a high data throughput.

Generally speaking, a data center is based on two main components: servers, responsible for information storage and processing, and the network (also termed Data Center Network (DCN)), which is responsible for the inter-server data delivery and thus the execution of the multi-server-based applications. Due to the ever decreasing costs of the silicon technology, manufacturers are able to double the number of transistors on their chips every 18 months (Moore’s Law), thus contributing to higher-performance machines with increasing processing power, as well as larger storage capacities for DC servers. As a matter of fact, Moore’s Law is not valid anymore when it comes to the networking fabric of a data center. For example, one of the network main parameters, the data throughput, is largely determined by the transistor speed and the limited number of connector pins available on the CMOS switching chips – characteristics for which Moore’s Law does not apply. Since most of the DCN components (switches, routers, etc) are electronic devices, their port-number limitations make the data center scalability impossible without an extensive deployment of a vast array of electronic CMOS switching chips, which subsequently results in complex switching fabrics and a huge number of interconnections. Furthermore, electrical copper cables are no longer suitable for data rates required by cloud applications, hence optical fiber cables (mostly multi-mode) have
been extensively deployed as the transmission medium of choice within the DCN. Obviously, due to the electrical nature of the modulating data signal and the optical nature of the carrier, a constant optical-electrical-optical (OEO) conversion needs to be performed at each stage, when data pass a certain switching point on its way to the destination server. This naturally results in extra delays associated with the processing time of the electrical switches. Moreover, the large number of intermediate (electrical) switching nodes consumes a huge amount of energy – a crucial parameter largely influencing the Data Center architecture and efficiency. Furthermore, it is rather expensive.

In this context, the deployment of optical switching capabilities has gained an increasing attention in the past couple of years, both from academia and industry [2]. Optical switching offers a lot of important advantages, among which are: no need for OEO conversion at each point of transition, a significantly reduced energy consumption compared to the traditional electrical-packet-switches (EPS), a much higher port density, which allows a simplification of the DCN architecture, and last but not least, the capability of transmitting multiple source signals on one single fiber and their consequent switching on one single port due to the wavelength division multiplexing (WDM) technology. The main drawback of optical switching, though, remains the relatively long switching setup time, which allows a simplification of the DCN architecture, and last but not least, the capability of transmitting multiple source signals on one single fiber and their consequent switching on one single port due to the wavelength division multiplexing (WDM) technology.

2 Traffic Modeling

The server generated data, as opposed to voice traffic, exhibit a highly variable nature in time, displayed by spurts of activity and traffic bursts, which can last from milliseconds to days, and look similarly independent of the time scale [8]. This property is called self-similarity and is characterized by the so-called Long Range Dependence (LRD), meaning that the generated data traffic strongly correlates at all time scales. A distinguishing peculiarity of LRD traffic is that the variation of the data rate around the mean value is very high. By contrast, the voice data traffic, which typically behaves as a Poisson process and a particular case of the Short Range Dependence (SRD) characteristic, exhibits a much lower variation around the mean value. Since LRD behaviour manifests itself as a bursty and highly variable activity in the data rate, which persists on different time scales, it becomes hard to predict how the data rate will alter over time [8, 9].

2.1 Mathematical Model

Since the LRD property describes the machine-originated data traffic more realistically than a Poisson process, its mathematical basis serves as a good foundation for the server packet generation model. One possible implementation of the LRD property is described by means of the so-called chaotic maps (Figure 1), defined by the function \( f = f(m_1, m_2, d) : I \rightarrow I \), in the unit interval \( I = [0, 1] \) as \( x_{n+1} = f(x_n) \) [8]:

\[
\begin{align*}
x_{n+1} &= \begin{cases} 
1 - x_n + (1 - d) \left( \frac{x_n}{d} \right)^{m_1}, & 0 \leq x_n \leq d, \\
1 - x_n - (1 - d) \left( \frac{1 - x_n}{1 - d} \right)^{m_2}, & d < x_n \leq 1,
\end{cases}
\end{align*}
\]

with \( 0 < d < 1 \) and the initial condition \( x_0 \).

![LRD traffic generation model according to (1). Blue dots in the interval \([0.5,1]\) represent packet generation instances; within \([0,0.5]\) no packet generation occurs.](image)

The parameter \( d \) serves as a decision threshold, meaning that if \( x_0 \) falls between 0 and \( d \), no packet is generated, while for \( d < x_0 \leq 1 \) packet generation occurs. In this case,
the mapping associated with the packet generation model has the form:

\[ y_n = \begin{cases} 
0 : 0 \leq x_n \leq d - \text{no packet generation}, \\
1 : d < x_n \leq 1 - \text{packet generation}.
\end{cases} \]  

(2)

The parameters \( m_1, m_2 \in ]1.5, 2] \) induce intermittency (alternation of traffic generation and non-generation). It is worth noticing that for values of \( 1 \leq m_1, m_2 \leq 1.5 \), the generated data traffic exhibits an SRD behavior (Poisson traffic), namely due to the exponential decay of the SRD queue length, as opposed to the power law decay specific to LRD sources. The intermittency nature of the mapping \( f \) introduces the memory effect in the digital output \( y_n \), which basically results in the long range pattern correlation of the generated data traffic [8].

### 2.2 Traffic Generation

Based on the LRD model, we simulate the traffic generation at the server level. In this context, it is worth emphasizing that the two intermittency parameters \( m_1 \) and \( m_2 \) are set to their maximum values: \( m_1 = m_2 = 2 \), in order to generate a highly variable and alternating data traffic, while keeping the same probabilities for packet generation and non-generation. Also, an important parameter, namely the generation threshold \( d \), is varied in the following in order to analyze the data generation and non-generation behavior. We set a generation time interval of \( 10^3 \) time-instances, in order to produce a more precise data sequence, at the same time preserving and visualizing the effects of LRD on the traffic generation itself.

Now, an important aspect which needs to be considered is the nature of the generated data traffic. LRD is characterized by highly variable spurts of generated data packets, or traffic bursts, which are short in time and highly variable in terms of data rate. These patterns are commonly addressed as *mice*. Mice are normally short connections with respect to the transmission time and the amount of data they carry, and have a higher occurrence frequency than other types of traffic flows [10]. A second type of LRD data stream occurring much more rarely in time, though carrying a much higher amount of data (compared to mice), and which can also last for longer periods of time, is the so-called *elephant*. By definition, a traffic flow is designated as an elephant based on both, its carried data volume and its persistence in time [11]. Mice and elephants occur with different frequencies, and their generation can easily be recognized in Figure 2. In this context, the main purpose of the hybrid DCN schemes is to differentiate between the two data streams, where the highly variable mice are routed through the EPS grid, while the longer and intensive elephants are switched through the established circuits of the OCS fabric. This distribution of traffic flows is due to the differing natures and purposes of EPS and OCS switching. As it is commonly known, packet switching is advantageous for short and variable traffic patterns, namely due to its very short switching time. By contrast, circuit switching offers dedicated paths between the source and destination nodes – a mechanism best suited for long-lasting and bulky traffic flows.

In order to differentiate between mice and elephants, we take into account the variation of the data rate as a function of time: \( R = f(t) \). That is, since the elephants tend to display a much more constant data rate over a long period of time, whereas mice’ data rate is highly variable within short time spans, we introduce a variation threshold, which basically defines whether a certain data stream behaves as a mouse or an elephant. It is worth emphasizing, though, that it is crucial to also consider the time span within which this alteration occurs. An example of such a threshold is a variation of \( \pm 5 \% \) of the server link throughput (i.e. a \( \pm 0.05 \times 10 \text{ Gbps} = \pm 0.5 \text{ Gbps data rate variation} \)), within two adjacent time instances \( T_n \text{ and } T_{n+1} \). This time interval is very relevant for the decision taking regarding switching path selection, since the previously generated data need to be stored in the cache memory, until the next information pattern will be produced for comparison. Furthermore, after the decision is taken, the data have to be stored in the cache until the optical circuit is set up to the final destination (through the OCS). As a result, as long as the data rate does not exceed a variation of \( \pm 5 \% \) within an interval of two adjacent time instances, the traffic flow can be considered an elephant: \( y_n - 0.5 \leq y_{n+1} - 0.5 \). Obviously, for a higher precision granularity, this threshold can be further lowered, or by contrast, increased, to offer a coarser decision

![Figure 2](image-url)  
**Figure 2** Server traffic generation: Interface utilization \( = f(t) \). Two traffic patterns, mice (red circle) and an elephant (green ellipse) occurring during data generation, are highlighted. The mice, which appear as traffic bursts, are packets with different destinations, routed through the EPS grid. The elephants, with a much longer duration and data amount carried, are dedicated to a singular receiver and thus routed through the OCS.
larger amount of data streams considered being elephants). In the following, the decision concerning choosing the electrical or optical paths will rely on this threshold value.

Based on the earlier introduced decision threshold and the server traffic generation simulated in Figure 2, it becomes straightforward to differentiate between mice and elephants traffic patterns (Figure 3).

3 DCN Architecture

In order to perform an architecture optimization and its subsequent evaluation, a distinct DCN structure has to be proposed. In contrast to the mega data centers, which can house 100,000+ servers, we base our analysis on a large data center, spanning around 32,000 servers. The servers are grouped into 1,000 racks, each containing 32 machines. Based on that, each server has a 10 Gbps NIC adapter, each connected to a ToR switch by means of an optical SFP+ interface. In the first base scenario, we assume a pure EPS-based Fat Tree Architecture, where each ToR is connected to the Aggregation switch by means of two active optical cables (AOC) with a QSFP28 interface and a throughput of up to 100 Gbps each, resulting in an oversubscription factor of $OF = 1:1.6$. The aggregation switches are further connected to the Core level by means of eight ports with an OSFP interface (currently in development) each, providing a data throughput of up to 400 Gbps/port. In the second, hybrid electro-optical Fat Tree architecture, an OCS intermediary level is introduced, with the ToR switches using one optical connection to the EPS aggregation switches, and a second connection routing the bulky traffic through the OCS. This architecture optimization automatically decreases the number of aggregation and core switches by a factor of 2, which will be discussed in more detail in Section 5. Obviously, the implementation of the hybrid scenario pursues the simplification of the DCN architecture by reducing the quantity, and hence the capital expenditure (CapEx) costs, as well as cutting the operation expenditures (OpEx) associated with the energy-hungry EPS devices. Moreover, the OCS energy efficiency not requiring OEO conversion, makes the hybrid scenario even more essential. The traditional and hybrid architectures can be seen in Figure 4.
The architectural details are summarized in Table I. We also consider an extended hybrid scenario in Table II, where data rates (link throughputs) between different levels are increased by a factor of 10.

**Table I** Base scenario. DCN Architecture components (pure EPS vs. hybrid electro-optical)

<table>
<thead>
<tr>
<th>Component</th>
<th>Interface/Throughput</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>SFP+/10 Gbps</td>
<td>32.000</td>
</tr>
<tr>
<td>ToR Switch</td>
<td>QSFP28/32x100 Gbps</td>
<td>1.000</td>
</tr>
<tr>
<td>Aggregation</td>
<td>QSFP28/100 Gbps</td>
<td>63/32</td>
</tr>
<tr>
<td>OCS</td>
<td>16x400 Gbps</td>
<td>32/16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>Interface/Throughput</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>QSFP28/100 Gbps</td>
<td>32.000</td>
</tr>
<tr>
<td>ToR Switch</td>
<td>QSFP28/32x100 Gbps</td>
<td>1.000</td>
</tr>
<tr>
<td>Aggregation</td>
<td>QSFP28/24x400 Gbps</td>
<td>94</td>
</tr>
<tr>
<td>OCS</td>
<td>48x400 Gbps</td>
<td>47</td>
</tr>
</tbody>
</table>

**Table II** Extended scenario. DCN Architecture components (hybrid electro-optical only)

<table>
<thead>
<tr>
<th>Component</th>
<th>Interface/Throughput</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>QSFP28/100 Gbps</td>
<td>32.000</td>
</tr>
<tr>
<td>ToR Switch</td>
<td>QSFP28/32x100 Gbps</td>
<td>1.000</td>
</tr>
<tr>
<td>Aggregation</td>
<td>QSFP28/24x400 Gbps</td>
<td>94</td>
</tr>
<tr>
<td>OCS</td>
<td>48x400 Gbps</td>
<td>47</td>
</tr>
</tbody>
</table>

4 Evaluation

Since this work primarily focuses on the Fat-Tree architecture, which is characterized by a hierarchical structure, starting with the Server level and followed by the ToR, aggregation and core levels respectively, we base our conducted evaluation on this order.

The server traffic generation model has been previously discussed and analyzed in more detail in Section 2.2. The LRD packet generation model proved its highly variable nature, where different types of traffic patterns have been exhibited. As a matter of fact, the behavior of the server level remains valid for both analyzed scenarios.

4.1 ToR Level

Starting with the ToR level, the term of data aggregation comes into play. In the context of DCN, data aggregation means gathering of the generated traffic through $N$ server interfaces and its further routing to the aggregation level using fewer interfaces. In this respect, the term oversubscription factor is used to indicate the ratio between the northbound and southbound throughputs. In our base scenario we consider a capacity of 320 Gbps southbound and a capacity of 200 Gbps northbound, which yield an oversubscription ratio of 1:1.6. Emerging from the term itself, the oversubscription ratio shows that the northbound interface is capable of handling less data than the southbound interface can generate. This is due to the fact that most often not all server ports are busy in sending data, which means idle states happen quite often, when little to no data traffic is generated. In fact, by using the earlier discussed data generation threshold, $d = 0.5$, we define the data generation and non-generation instances as equiprobable. Nevertheless, when talking about oversubscription ratio, there is always a port blocking probability, $P_B$. In terms of DCN performance, $P_B$ becomes a crucial parameter influencing the network design and its subsequent architecture.

Another important aspect considered in this work, and also having an impact on the $P_B$ value, is the intra-rack server communication. In particular cases, the data traffic generated by servers is simply switched locally by the ToR, thus never reaching the EPS or OCS aggregation layers. Such data streams make up to 13% of the total rack-generated data [12], hence it becomes crucial to take this particularity into consideration for analysis. This percentage is of course application-based, depending on whether the rack servers run a service together, or the application is distributed over multiple racks.

In order to determine the blocking probability of the ToR, a data generation and aggregation simulation has been deployed. In our simulations 1,000 different loops have been performed. In the following figures also the 95% confidence intervals are depicted. The blocking probability has been simulated as a function of packet generation intensity: $P_B = f(d)$, with a subsequent estimation of the confidence interval (Figure 5).

Another important aspect worth considering was the bandwidth mean utilization $\eta$, which again was measured as a function of packet generation intensity at the rack level (Figure 5). At $d = 0.5$ the mean utilization is roughly 45%. In this respect, due to the highly variable nature of LRD traffic, despite only 45% of the uplink throughput is on average consumed, there is still a 2% probability that the northbound ports will experience a congestion. This limitation makes it crucial to provision 200 Gbps throughput in the northbound link.

When moving towards the hybrid electro-optical architecture, the emphasis has been put on the data streams’
analysis. As previously discussed in Section 2.2, the hybrid scheme uses the EPS grid for mice switching, whereas the stable optical circuits set through the OCS network are dedicated to long-lasting and bulky elephants. In this respect, we distributed the two traffic patterns on to the two northbound ToR interfaces each. Hence, both the mice and the elephants have an equal throughput for data transmission. In order to assess the scheme’s performance, the interface mean utilization, $\eta$, and the port blocking probability, $P_B$, have been simulated (Figures 6 and 7).

In order to remove this performance bottleneck, a second northbound link of 100 Gbps dedicated to elephant streams has been added. In this context, the main advantage of the hybrid scheme is that with the addition of such a link, the architectural complexity does not escalate, namely due to the possibility of multiplexing two elephant-dedicated northbounds into one single OCS port [13].

As expected, the link to the OCS reaches rapidly a high port blocking probability $P_B$ (Figure 7) due to the huge amount of generated and transmitted data, while the interface to the aggregation switch displays a near-to-zero blocking probability. It is important to emphasize that for low values of $d$, i.e. high packet generation intensity, the elephants prevail over the short mice, whilst when $d$ starts increasing, mice generation becomes dominant. Still, due to the throughput and the corresponding data type imbalance, even for $d = 0.5$ the blocking probability of 46.7% of the elephants interface is unacceptable in terms of transmission performance (Figure 8).

As it can be seen from Figure 9, after the addition of the second link, $P_B$ has dropped significantly, up to the point where it barely reaches 0.5% for both, mice and elephants (at the operation point $d = 0.5$). Another important aspect is that the mice-dedicated 100 Gbps northbound link is certainly more than enough to ensure an almost blocking-free transmission, thus raising the question whether its throughput is efficiently used. Nevertheless, downgrading it to the lower neighboring interface QSFP with 40 Gbps is not a solution either, since the mice blocking probability in this case rises dramatically (up to 60-70%), plus the market available ToR switches do not offer such a large range of interfaces (SFP+, QSFP and QSFP28) on one single piece of equipment.
4.2 Aggregation and Core Levels

Until now, the analysis has been conducted at the ToR level. In the context of hybrid architecture, the most important decisions regarding data traffic type and its subsequent routing to the proper switching grid, is mainly conducted at the ToR layer. In this regard, it has to be properly decided whether a data stream is suited for optical or electrical switching. Moreover, the incoming data need to be buffered at least for one data frame interval, in order for the decision upon data rate variation and the corresponding data type to be taken. Furthermore, delays due to OCS switching setup time need to be bridged.

After the ToR traffic aggregation procedure, the next level is the Aggregation layer. In our considered scenario, aggregation switches ensure a 3.2 Tbps (32x100 Gbps) southbound throughput, and a 3.2 Tbps (8x400 Gbps) northbound throughput. The northbound bandwidth has been determined after a number of simulations and adjusting its value accordingly, in order to ensure zero blocking probability. In the base scenario, the aggregation switch processes both data types, as opposed to the hybrid scenario, where only mice are transmitted. This leads to a visible variation in the data rates, and a much more bursty and variable nature of the data traffic (Figure 10).

Since the hybrid architecture preserves only one single connection between the ToR and the EPS grid, this cuts the amount of aggregation switches by a factor of 2, compared to the traditional base scenario. As the aggregation switches are further connected to the core layer using the OSFP 400 Gbps interfaces, the core southbound throughput displays similarly a $P_B = 0\%$. In our proposal the core switches are equipped with 16 OSFP interfaces, ensuring a bandwidth of up to 6.4 Tbps. Once the amount of aggregation switches reduces by half, the same applies to the core switches.

Similarly with the intra-rack local data, part of the aggregated traffic never reaches the core layer, instead is being switched to other ToRs by aggregation switches. According to Cisco [12], up to 77 % of the aggregated data traffic remains within the DCN (excluding the intrarack local data), and it becomes a challenge to differenti-
relative to a reference piece of equipment. That is, the reference price of pluggable components is computed relative to the SFP+ price. On the other hand, the switching components’ prices are calculated relative to the basic switching unit, i.e. the ToR switch with two northbound interfaces. Since we focus our evaluation on the data center network architecture, the servers’ cost will not be considered in this analysis. The relative costs per unit of equipment are presented in Table III (in price units).

### Table III
Relative normalized costs of single unit components, and their corresponding energy consumption (base scenario).

<table>
<thead>
<tr>
<th>Components</th>
<th>Cost, p.u.</th>
<th>Energy, W</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToR Switch (2/3 ports)</td>
<td>1/1.09</td>
<td>660</td>
</tr>
<tr>
<td>Aggregation Switch</td>
<td>1.48</td>
<td>700</td>
</tr>
<tr>
<td>Core Switch</td>
<td>1.48</td>
<td>700</td>
</tr>
<tr>
<td>OCS</td>
<td>8.7</td>
<td>100</td>
</tr>
</tbody>
</table>

Table IV: Total relative CapEx (total cost) and OpEx (total energy) consumptions (base scenario).

<table>
<thead>
<tr>
<th>Components</th>
<th>Electric</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cost, p.u.</td>
<td>Total energy, kW</td>
<td>Total cost, p.u.</td>
</tr>
<tr>
<td>ToR Switch</td>
<td>1000</td>
<td>660</td>
</tr>
<tr>
<td>Aggregation Switch</td>
<td>93.24</td>
<td>44.1</td>
</tr>
<tr>
<td>Core Switch</td>
<td>47.36</td>
<td>22.4</td>
</tr>
<tr>
<td>OCS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Σ, Switches</td>
<td>1140.6</td>
<td>726.5</td>
</tr>
<tr>
<td>SFP+</td>
<td>32000</td>
<td>32</td>
</tr>
<tr>
<td>QSFP28</td>
<td>40160</td>
<td>13.25</td>
</tr>
<tr>
<td>OSFP</td>
<td>50800</td>
<td>12.19</td>
</tr>
<tr>
<td>Σ, Pluggables</td>
<td>122960</td>
<td>57.44</td>
</tr>
</tbody>
</table>

As it can be concluded from Table IV, the hybrid scheme brings an improvement in CapEx, due to a much smaller quantity of pluggables (20% less), which represent a significant percentage of the DCN cost. Moreover, the hybrid scheme yields a lower total energy consumption (OpEx) as opposed to the pure electrical architecture.

### 6 Conclusions

The main aim of this work was to analyze how hybrid electro-optical switching impacts the performance and costs of a traditional Fat-Tree data center network. Special attention has been paid to traffic modeling, with emphasis on differentiating the different data patterns of a server (mice and elephants). Since mice exhibit a highly variable and bursty nature, they behave very differently than the more stable and longer lasting elephants. By sending the latter long data streams through the optical circuit switched network and the former short traffic bursts through the electrical packet switches, data centers can become more efficient and significantly cheaper. Furthermore, the electrical energy consumption can be reduced, leading to lower OpEx. In future DCNs, optical switching will certainly become a necessity due to the ever increasing interface speeds and purely optical intra-datacenter networks are at the horizon.

### 7 References


