Modeling and Evaluation of a Data Center Sovereignty

Shakthivelu Janardhanan
Chair of Communication Networks
Technical University of Munich
Munich, Germany
shakthivelu.janardhanan@tum.de

Carmen Mas-Machuca
Chair of Communication Networks
Technical University of Munich
Munich, Germany
cmas@tum.de

***Abstract*—Technology Sovereignty aims at protecting the interests of the consumers belonging to a sovereign state. The new laws on network sovereignty monitor and govern networks inside a state. They protect the constitutional rights of citizens and ensure data security. However, they do not focus on the infrastructure and hardware associated with these networks. Ensuring a robust end-to-end network fuels the need for reliable hardware manufacturers and an appropriate network architecture that can handle multiple failures efficiently. To understand technological sovereignty in communication networks, the Data Center Network (DCN) is targeted as it is a critical part of the digital society. Data centers are dedicated physical facilities that act as storage houses for large amounts of data. Though several fault-tolerance studies have been performed in DCNs, none have studied the role of hardware manufacturers in DCN Sovereignty. The unavailability of components from a manufacturer can lead to multiple failures. So, it is necessary to build a sovereign DCN without creating a dependency on the manufacturer(s). In this work, to evaluate the sovereignty of a DCN, (i) Multiple failure scenarios depending on manufacturer reliability are evaluated, and (ii) Design guidelines on how to choose the number of hardware manufacturers and how to arrange them in the DCN topology are presented.**

***Index Terms*—sovereignty, reliability, data center**

I. Introduction

Growing uncertainties concerning trade conflicts and economic crises have raised questions regarding the interdependence of economies. The rise or fall of a state's economy can influence another state. Without a well-defined limit on this influence, it is impossible to ensure stability. With the growing demands in communication networks, the need for establishing 'Strategic Autonomy' has become a topic of debate [1], [2]. Core technologies need to remain unaffected by economic and geopolitical influence. But, building an entirely indigenous technology without any dependency on foreign solutions is very difficult. Hence, it is wiser to ensure that the technology does not have any detrimental large-scale structural dependencies on a foreign solution.

Data centers have become the pillars of the Information Age. They play a crucial role in the growth of a state's digital economy. In this study, DCNs are considered because they are (i) Continuously evolving, and (ii) Relevant in the foreseeable future. In a DCN, having backup paths is a common solution to ensure optimal performance even under stress. But, studies like [3] show that the redundancy in a DCN is not fully effective. So, it is required to make conscious planning decisions on the placement of the hardware and their interconnections based on the reliability of their manufacturers.

This work looks at the worst-case scenarios possible, i.e., the effect of massive failures with respect to the manufacturer's reliability in a DCN. Massive failures may occur due to unavailability of a particular component, unavailability of an element inside a component, misbehavior of a component, incompatibility due to firmware upgrades, or many other reasons. This is where the placement of these components from different manufacturers comes into play. In short, this research focuses on analyzing the impact of,

(P1) choosing the number of manufacturers (N_m),
(P2) setting an upper bound (U_{mi}) on the number of components from each manufacturer, and
(P3) their placement in the DCN topology,

on the DCN robustness in terms of connectivity, max-flow, and load on the links. This research addresses multiple fault-tolerance scenarios based on the manufacturer's reliability.

This paper is organized as follows: Section II discusses the related work. Section III describes how the DCN has been simulated. Section IV describes the implementation specifics in the simulation. Section V summarizes the findings of this work. Section VI concludes this paper.

II. Related Work

Digital sovereignty is the ability to develop and provide or outsource the technology [2]. Outsourcing may lead to a dependency on one or a small set of providers. In this case, the reliability of the providers dictates the success of that technology. In this paper, technological sovereignty in data centers is studied.

Many studies like [3]–[5] have gathered statistics from data centers and have shortlisted the most prominent causes. The major causes include hardware faults, maintenance issues, firmware bugs, misconfigurations, accidents, and capacity planning. [4] reports that the reasons for 29% of intra-DCN issues at Facebook from 2011 to 2018 are undetermined.

The major findings from these studies are as follows. Software issues are more common than hardware failures [4], [5]. Software issues have smaller downtime than hardware failures [3]. Failures in Top of the Rack switches account for the majority of downtime [3], [4]. Fabric networks (fat tree-based networks) have less severe issues than clusters (Leaf-Spine based networks) [4]. Bigger networks have longer time-to-repair [4]. Though the network has sufficient backup paths between any communicating pair, the failure of devices still affects the performance considerably [3]. In [6], the authors propose estimating the expected time-to-failure of switches and taking preventive actions. They employ the use of a Kaplan Meier survival estimator and a Proportional Hazard model to identify the effect of the switch's hardware/software features on its time-to-fail estimation.

These studies work on identifying the issues and their causes. However, they do not consider the impact of the manufacturers of the components, nor do they enclose any information regarding the arrangement of the components. They do not consider the possibility of a mass failure of components. In this paper, simultaneous failure of several devices is considered for the sovereignty analysis.

III. DCN Modeling

The major components of a DCN relevant to this study are servers, Top of the Rack switches (ToR), aggregate switches, and core switches. A Pod is a basic entity in a DCN that improves scalability. It consists of a fixed number of ToR and aggregate switches. Throughout this research, Top of Rack architecture is considered. DCN Modeling can be divided into three sections - macro parameters, micro parameters, and input options.

1) Macro parameters: They are concerned with the overall characteristics of the simulator. They include simulation duration (T_S), number of flows per server per second (N_{SS}), oversubscription (R_{OS}), intra to inter-rack traffic ratio (R_{rack}), and intra to inter-DCN traffic ratio (R_{DCN}). Throughout this work, T_S has been fixed at $1s$. In [7] written in 2015, N_{SS} varies from 200 to 500. To match today's traffic expectation, N_{SS} requires extrapolation based on [8]. [8] shows up to a $1.6\times$ increase in the workload per server between 2015 and 2020. Using this factor, in this work, N_{SS} has been set to 800. The R_{OS} of a switch is the ratio of the total bandwidth of all southbound ports to all northbound ports [9]. For a large DCN with over 32K servers, maintaining a non-blocking network with an R_{OS} of $1:1$ is not feasible. The R_{OS} varies from $2.5:1$ to even $240:1$ [10]. Typically, the R_{OS} tends to be smaller to accommodate large bursts of traffic. In this research, each ToR switch has an R_{OS} of $3:1$, while each aggregate switch has an R_{OS} of $2:1$. Only the Facebook 4-post topology, discussed in Section III-3, is an exception to this rule [11]. Here, the ToR switch has an R_{OS} of $10:1$, while the aggregate switch has an R_{OS} of $3:1$. The majority of flows from a rack stay within the rack [12]. So, the R_{rack} has been fixed at $70\%:30\%$. As per [13], the inter-DCN traffic is considered as a 'special rack' and the R_{DCN} is fixed at $90\%:10\%$.

2) Micro parameters: They control the flow characteristics and traffic management. They include the flow size, Traffic Matrix (TM), and link utilization.

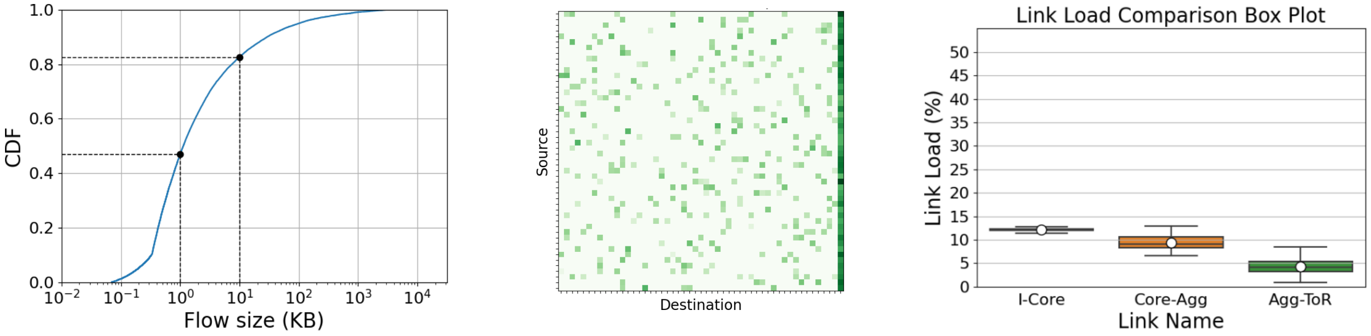
Most of the flows (80%) are small in size, smaller than $10KB$ [14]. 95% of the flows are smaller than $1MB$ and 99% of the flows are smaller than $100MB$ [14]. Most of the bytes are associated with the top 10% of the large flows [7], [15]. Most of the flows (80%) have a short duration, lesser than $10s$ [7]. This work does not use the flow duration. Only the flow size is considered. To model the flow size, a modified Pareto distribution is used to achieve the required curve as shown in the literature. From Fig. 1a from the simulator, nearly 45% of the flows are smaller than $1KB$ while a little over 80% of the flows are smaller than $10KB$.

The TM shows the communicating pairs of servers in the DCN. The TMs between ToRs are sparse [13]. Fig. 1b shows a portion of the TM generated in the simulator for a small-sized fat tree topology. Each grid in the heat map corresponds to several flows with source as the ToR in the Y-axis and destination as the ToR in the X-axis. The last column is the inter-DC connection. Thus, the TM in the simulator is sparse. The intra-rack traffic is not shown in this figure.

The link utilization at the core layer is greater than the aggregation layer, which is in turn greater than the ToR layer [12]. The mean ToR-Server link utilization is lesser than 1% [7]. 99% of all the links are usually lesser than 10% loaded [7]. The median ToR-Agg link utilization varies between $10-20\%$ [7]. The busiest 5% of the links are 23–46$\%$ loaded [7]. This work considers only the link load. The link load is explained in IV-1. Employing a particular architecture for routing falls outside the scope of this work. So, a common ECMP algorithm is used for routing the flows in this work.

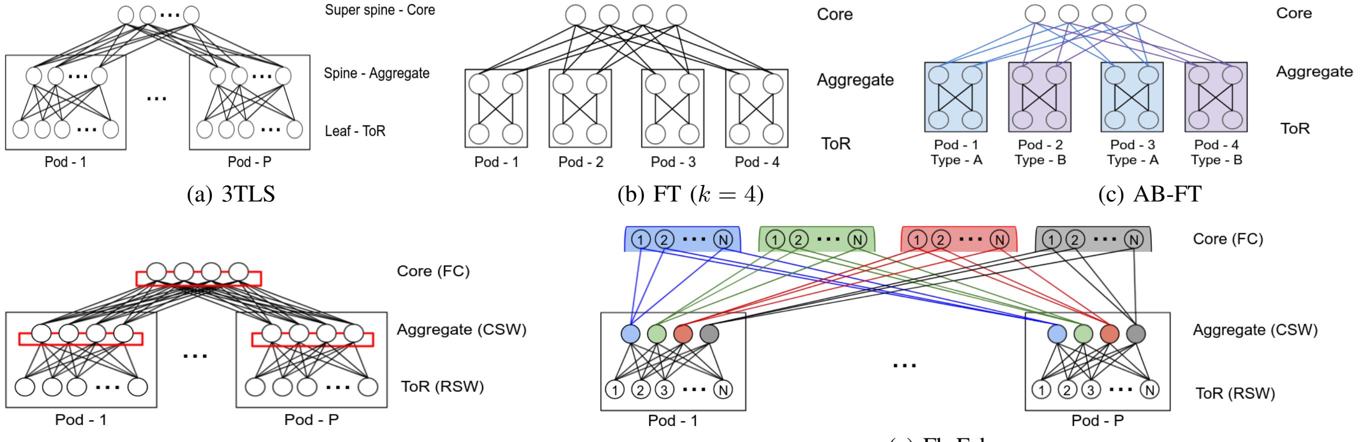
3) Input Options: The major input options - topology, size, N_m, and arrangement are shown in Table I. There are a total of 320 input combinations possible.

The topologies considered in this work are summarized in [11] and [16]. The Leaf-Spine topology consists of a lower layer of switches called leaf switches and an upper layer of switches called spine switches in each pod. Every leaf switch connects to every spine switch. This can be extended to another layer consisting of super spine switches, referred to as a 3 Tier Leaf-Spine topology (3TLS) as seen in Fig. 2a. The Fat Tree topology (FT) uses a Clos-network for scalability as seen in Fig. 2b. It uses commodity switches in all the layers. Each core switch is connected to one aggregate switch in every pod. The number of switches in a pod, number of core switches, and interconnections are all calculated based on the number of pods (k). The robustness of FT is improved in the AB-Fat Tree topology (AB-FT) by skewing the FT as seen in Fig. 2c. There are two types of pods alternately arranged in the AB-FT. In type-A pods, each aggregate switch is connected to consecutive core switches. In type-B pods, each aggregate switch is connected to the core layer in steps of fixed length. The Facebook 4-Post topology (Fb-4P) shown in Fig. 2d is a modified 3TLS topology. The number of aggregate switches (cluster switches) per pod (n_A) and the number of



(a) Flow Size-CDF (b) ToR-TM for flows (c) Link Load box plot

Fig. 1: Images from the simulator



(a) 3TLS (b) FT ($k = 4$) (c) AB-FT

(d) Fb-4P (e) Fb-Fab

Fig. 2: Topologies

core switches (fat cats) (N_C) is always 4. The structure and link capacity are exactly as suggested in [11]. The Facebook Fabric topology (Fb-Fab) shown in Fig. 2e also has 4 aggregate switches per pod. There are 4 planes in the core layer. Each aggregate switch is connected to all the core switches in one plane. Each core switch is connected to only one aggregate switch in every pod.

Sizes ranging from a thousand to a hundred thousand servers were considered in this paper. Depending on the topology and size, the infrastructure followed for the various topologies are given in Table II. The notation used in this table is explained in Table III.

The DCNs do not operate with only one manufacturer as it may become a single point of failure. Though the simulator supports any value of N_m, this work considers four options - 2, 3, 4, and 5.

The four options for arranging the components from the manufacturers- Random (A_{Ran}), Left-Right (A_{LR}), Left-Right Sequential (A_{LRS}), and Pod-wise (A_{Pod}), are illustrated in Fig. 3. Each circle is a component from the same layer. Each rectangle is a pod. Each color is a different manufacturer. Let the N_m be 3. This is denoted by blue, purple, and yellow. Unless specified as heterogeneous distribution, an approximately equal number of components are purchased from each manufacturer. In the A_{Ran}, the components are randomly placed in the DCN. The output for the A_{Ran} is calculated as an average of multiple simulations to ensure fairness. In the A_{LR}, the components in a layer are divided into a number of sections, equal to the N_m. So, each section corresponds to one manufacturer. In the A_{LRS}, the components in a layer are arranged sequentially, for example, blue-purple-yellow. This sequence is repeated. In the A_{Pod}, all the components in one pod are from the same manufacturer.

Parameters	**Options**
Topology	3TLS, FT, ABFT, Fb-4P, Fb-Fab
Size	Small (1K Servers), Medium (32K Servers), Large (64K Servers), Mega (100K Servers)
N_m	2,3,4,5
Arrangement	Random, Left-Right, Left-Right Sequential, Pod-Wise

TABLE I: Input Parameters

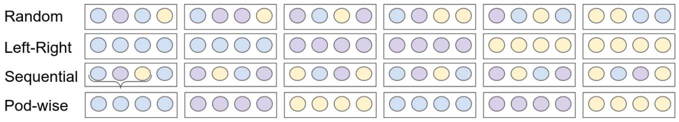


Fig. 3: Examples of the different arrangements

IV. Sovereignty Analysis

Throughout this work, the servers are connected to only one ToR. So, the failure of a ToR means all the flows to and

Topology	Size	N_S	n_S	n_T	n_A	N_P	N_T	N_A	N_C	L_{ST}	L_{TA}	L_{AC}	L_{AI}	L_{CI}	R_A	R_C
3TLS	Small	~1K	48	22	6	1	22	6	0	10	40	-	100	-	-	-
	Medium	~32K	96	30	8	11	330	88	6	10	40	100	-	400	-	-
	Large	~64K	96	30	8	22	660	176	6	10	40	100	-	800	-	-
	Mega	~100K	96	30	8	35	1050	280	6	10	40	100	-	800	-	-
FT, ABFT	Small	~1K	20	5	2	10	50	20	4	10	40	40	-	40	-	-
	Medium	~32K	64	16	5	32	512	160	40	10	40	40	-	40	-	-
	Large	~64K	80	20	6	40	800	240	66	10	40	40	-	100	-	-
	Mega	~100K	92	23	7	46	1058	322	84	10	40	40	-	100	-	-
Fb-4P	Small	~1K	44	23	4	1	23	4	0	1	10	-	40	-	80	160
	Medium	~32K	44	48	4	15	720	60	4	1	10	40	-	400	80	160
	Large	~64K	44	48	4	31	1488	124	4	1	10	40	-	800	80	160
	Mega	~100K	44	48	4	48	2304	192	4	1	10	40	-	800	80	160
Fb-Fab	Small	~1K	48	21	4	1	21	4	0	10	40	-	40	-	80	160
	Medium	~32K	48	48	4	14	672	56	96	10	40	40	-	40	80	160
	Large	~64K	48	48	4	28	1344	112	96	10	40	40	-	80	80	160
	Mega	~100K	48	48	4	44	2112	176	96	10	40	40	-	100	80	160

TABLE II: Infrastructure followed for all topologies and sizes

N_S	Total no.of servers	n_S	No.of servers per ToR
N_T	Total no.of ToRs	n_T	No.of ToRs per pod
N_A	Total no.of aggregate switches	n_A	No.of aggregate switches per pod
N_C	Total no.of core switches	N_P	No.of pods
L_{ST}	Server-ToR link capacity (Gbps)	L_{TA}	ToR-Aggregate link capacity (Gbps)
L_{AC}	Aggregate-Core link capacity (Gbps)	L_{AI}	Aggregate-Inter-DC link capacity (Gbps)
L_{CI}	Core-Inter-DC link capacity (Gbps)	R_A	Aggregate ring capacity (Gbps)
R_C	Core ring capacity (Gbps)		

TABLE III: Notation used in this work

from that rack have failed. This loss of data occurs uniformly irrespective of the arrangement. So, it is not meaningful to consider the failures of servers and ToRs in this analysis. In this section, the output metrics are discussed, the notation for a failure scenario is explained, the failure scenarios are modeled, the different approaches for sovereignty analysis are explained, and the simulation flow is shown.

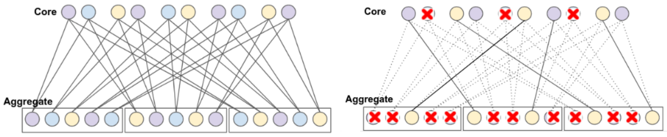
1) Output Metrics: The major output metrics considered in this research are - Connectivity, Max-flow, and Link Load.

The connectivity refers to the existence of a path between a pair of ToRs. It is a measure of the robustness of the topology in multiple failure scenarios. However, it does not guarantee successful traffic. Though a path exists, data can be lost due to overloaded links or overloaded CPUs of the switches in the path. The connectivity is still a valid output metric because the operator knows that the DCN can survive the majority of the traffic at lower rates. Connectivity Percentage (Z_C) is the ratio of the number of communicating pairs of ToRs for which a path exists to the total number of communicating pairs of ToRs. Z_C is plotted as heat maps for comparison between the various input configurations.

The minimum-cut separates the source and destination into disjoint graphs while minimizing the total weight on the edges that are being cut. The Max-Flow Min-Cut theorem states that the maximum flow between two nodes in a network is equal to the accumulated spare capacity of the edges in the minimum-cut. When there are no failures, the max-flow (z_f) is highest. As failures are introduced, the z_f reduces. The z_f reaches zero when no flow can be routed between the source and destination. If the available z_f is very small, that link has a greater chance of being overloaded. In this work, the z_f is calculated for every communicating pair of ToRs in the DCN. The average max-flow (Z_F) is the numerical mean of all the individual max-flows. The Z_F is plotted as heat maps for comparison between the various input configurations.

The link load is the sum of the sizes of the flows that use that link. The link load (%) is the percentage ratio of load on the link to the capacity of the link. A box plot of the link load is given in Fig. 1c. The link utilization is not considered as the simulations do not consider the rate of the flow. Only the size of the flow is considered. Since this paper is only concerned with the connectivity and the available bandwidth for the flows, the link load is in itself a fair metric for this analysis.

2) Notation for failure scenarios: The manufacturers are numbered from 0 through $N_m - 1$. T, A, and C denote the ToR, aggregate and core layers respectively. A combination of the letters like $[A, C]$ denotes both the layers- aggregate and core. The failure scenarios are written in the format $\{manufacturer\_id : [failed\_components]\}$. The set of all failure scenarios is given by F. For example, consider the failure scenario: $\{0:[A],\ 1:[A,C],\ 2:[\,]\}$ and the topology shown in Fig. 4a.



(a) No failures

(b) $\{0:[A],\ 1:[A,C],\ 2:[\,]\}$

Fig. 4: Example of a failure scenario

Let each color denote different manufacturers. In this example, manufacturer-0's aggregate switches have failed. So, all the purple nodes in the aggregate layer have been removed. Manufacturer-1's aggregate and core switches have failed. So, all the blue nodes in the aggregate and core layers have been removed. Manufacturer-2 has no failures. So, all the yellow nodes are present. After failure scenario, the topology looks like Fig. 4b. When a node fails, the edges associated with the node can not be used.

3) Failure scenarios: The various combinations of failures that are possible depend on the N_m. The possibilities for two and three manufacturers are shown in Table IV. The same logic is extended for any value of N_m. Simultaneous failure of all aggregate and/or core switch manufacturers is not considered because the results from this scenario will not change for different arrangements. In general, for every combination of topology, size, and arrangement, the number of possibilities is given by, $|F| = 1 + ((N_m - 1) \times 3)$. The no-failure scenario is also included as a benchmark.

N_m	$\|F\|$	F
2	4	$\{0:[\ ],\ 1:[\ ]\}$; $\{0:[A],\ 1:[\ ]\}$; $\{0:[C],\ 1:[\ ]\}$; $\{0:[A,C],\ 1:[\,]\}$
3	7	$\{0:[\ ],\ 1:[\ ],\ 2:[\ ]\}$; $\{0:[A],\ 1:[\ ],\ 2:[\ ]\}$; $\{0:[A],\ 1:[A],\ 2:[\ ]\}$; $\{0:[C],\ 1:[\ ],\ 2:[\ ]\}$; $\{0:[C],\ 1:[C],\ 2:[\ ]\}$; $\{0:[A,C],\ 1:[\ ],\ 2:[\ ]\}$; $\{0:[A,C],\ 1:[A,C],\ 2:[\,]\}$

TABLE IV: List of failure scenarios possible for $N_m = 2, 3$

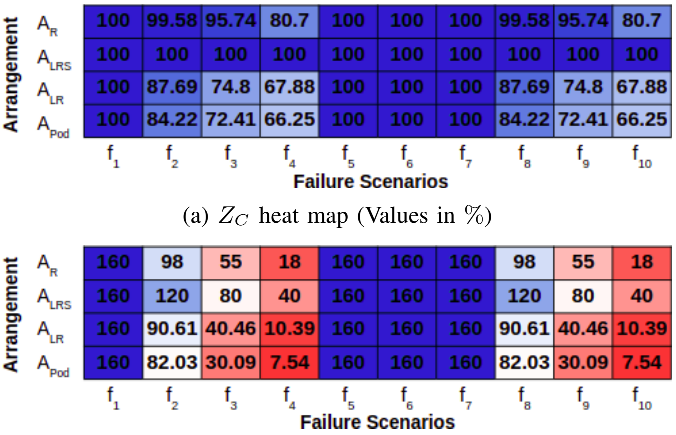
4) Analyses: There are three different analyses in this work-homogeneous, heterogeneous, and robustness surfaces.

The homogeneous analysis answers the problem statements (P1) and (P3). In this analysis, all manufacturers have an equal number of components in the DCN. The simulations are run for the various input configurations as shown in the pseudo-code 1. For each topology, size, and N_m, the output heat maps are generated with F along the X-axis and the arrangements along the Y-axis. The goal is to find the N_m and arrangement that suits the DCN requirements. Consider the example of a medium-sized Fb-Fab with $N_m = 4$. Fig. 5 shows the heat maps for Z_C and Z_F respectively. Along the X-axis, ten failure scenarios ($\{f_1, f_2, ..f_{10}\} \in F$) corresponding to $N_m = 4$ are considered. These failure scenarios are generated according to Table IV. These Along the Y-axis, the arrangements are considered. Each grid in Fig. 5a represents the Z_C for the corresponding failure scenario. Each grid in Fig. 5b represents the Z_F for the corresponding failure scenario in GB. The heat map coloring goes from red to blue, red being the worst and blue being the best.

The heterogeneous analysis answers the problem statements (P1) and (P2). The DCN operator would not buy an equal number of switches from each manufacturer. Let N_{mi} be the i^{th} manufacturer. Let N_{Ci} be the percentage of components from N_{mi}. Then, the upper bound for N_{Ci}, is given by U_{mi}, where $N_{Ci} \leq U_{mi}$. Here, one manufacturer (N_{m1}) with a varying number of components in the DCN is considered, i.e., $\{0\%, 10\%, 20\%, .., 70\%\} \in N_{C1}$. The simulations are run similar to the pseudo-code 1, but N_m is fixed at 2, because the focus is on U_{m1} only. For every simulation, all components from N_{m1} are failed. Consider the example of a medium-sized Fb-Fab shown in Fig. 6. In the output heat maps, the X-axis denotes N_{C1}. The first column is filled with maximum values because there are no failures. In the second column, N_{C1} corresponds to 10% of all components, and so on. Only the failures in the aggregate and core layers are considered. This analysis is expected to find the percentage of failing components that can be tolerated by the DCN at present and after some operational time without hardware upgrades. Based on assumptions made from traffic evolution studies, this paper presents a competent case study.

Algorithm 1: Simulation flow - Homogeneous analysis

```
Macro and micro parameters are loaded.
foreach topology do
    foreach size do
        foreach $n_m \in \{2, 3, 4, 5\}$ do
            foreach arrangement do
                foreach $f \in F$ do
                    Generate topology as a graph.
                    Generate flows with source,
                      destination and size.
                    Generate the TM.
                    Perform ECMP routing.
                    Plot macro and micro parameter
                      distributions for verification.
                    Save $Z_C$ and $Z_F$.
                end
            end
            Generate heat map for $Z_C$ and $Z_F$.
        end
    end
end
```

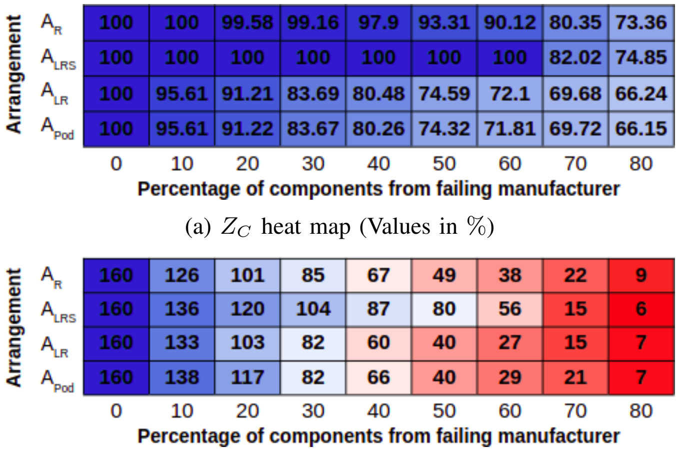


(a) Z_C heat map (Values in %)

(b) Z_F heat map (Values in GB)

Fig. 5: Homogeneous distribution: Fb-Fab, medium, $N_m = 4$

In any topology, when there are no failures, the z_f is the product of the number of uplinks per ToR and capacity of the uplink (C_{link}). Let T_{link} denote the traffic on the link. Let T_{acc} denote the accumulated traffic on all the uplinks of the ToR. As per Eq. 1, the link load (L_{link}) is the ratio of traffic on the link to the capacity of the link. In terms of the z_f, the L_{link} can be accumulated on all the uplinks of the ToR and compared against the accumulated capacity on all the uplinks (C_{acc}). But, C_{acc} is the available z_f. So, the total load on all uplinks of a ToR denoted by L_{acc} is given by the ratio of



(a) Z_C heat map (Values in %)

(b) Z_F heat map (Values in GB)

Fig. 6: Heterogeneous distribution: Fb-Fab, medium

accumulated traffic on all uplinks to the available max-flow.

$$L_{link} = \frac{T_{link}}{C_{link}}, \quad L_{acc} = \frac{T_{acc}}{C_{acc}} \tag{1}$$

$$\because C_{acc} = z_f, \quad L_{acc} = \frac{T_{acc}}{z_f} \tag{2}$$

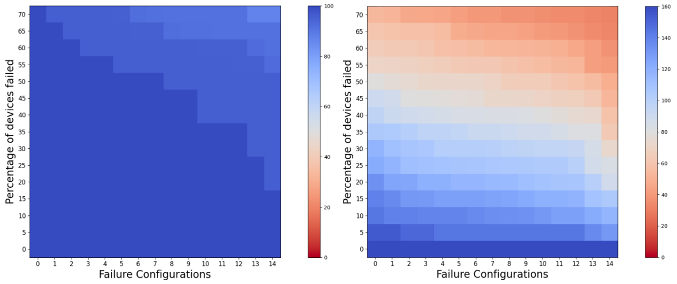
From Section III-2, the links that are heavily utilized have around twice the load of the other links. Optimistically, the other links can also handle twice their actual load. However, this statement is not applicable for the links that are already heavily loaded. So, realistically, assume that these links can handle about $1.5\times$ their loads without any congestion. So, the L_{acc} is increased by a factor of 1.5. When there are failures, the z_f reduces. The key is to find out how much the z_f can reduce before the links are overloaded, in this case, before the accumulated link load exceeds $1.5 \times L_{acc}$. Let x be the factor by which the z_f reduces under the failure scenario. Then, Eq. 2 is modified as,

$$1.5 \times L_{acc} = \frac{T_{acc}}{z_f \times x} \implies 1.5 = \frac{1}{x} \implies x = 0.66 \tag{3}$$

Thus, the z_f can reduce to 66% of its maximum value without traffic loss. However, this is only for the present traffic. In the future, the workload per server will increase, meaning the traffic on the links will increase.

Now, consider an example where the operator decides to build a data center in the year 2022 and does not want to make any hardware upgrades for 5 years. In the future, the incoming traffic will increase [8]. From Fig. 5 in [17], a $2.1\times$ increase in traffic is expected for the period 2022-2027. However, the workload on each server will not be doubled because the number of data centers in the world is also increasing [8]. Considering all this, a reasonable assumption of a $1.15\times$ increase in load per server can be made. So for the future, the z_f that can still tolerate failures is calculated by modifying the Eq. 3 as follows.

$$1.5 \times L_{acc} = \frac{T_{acc} \times 1.15}{z_f \times x} \implies x = 0.76 \tag{4}$$



(a) Successful Connectivity

(b) Average max-flow

Fig. 7: RS: Fb-Fab, medium, A_{LRS}, $N_m = 4$

Thus, the z_f can reduce to 76% of its maximum value without traffic loss. Since Z_F is the numerical mean of individual z_f values, the same conditions apply. This analysis can be extended to any other time period also.

The Robustness Surface (RS) [18] is an analytic tool that allows the summation and comparison of multiple output metrics with weights. The RS enables the visual comparison of multiple networks. For example, the medium-sized Fb-Fab with 4 manufacturers is considered in Fig. 7. Along the X-axis, different failure configurations are present. Along Y-axis, the percentage of devices that have failed is arranged. Failure configuration is the realization of failures in a particular order. Each configuration follows a different order.

Here, the components are purchased homogeneously with different failure rates. For example, N_{m1} has $2\times$ the failure rate of N_{m0}. N_{m2} has $3\times$ the failure rate of N_{m0}, and so on. This is an ideal scenario. If the operator buys more components from the least reliable manufacturer, then the DCN is more likely to lose traffic in the case of failures. For the 1^{st} row from the bottom, there are no failures. From the 2^{nd} row, for every row, 5% of the components are to be failed incrementally. For each simulation, the Z_C and Z_F are noted in two different matrices. This differs from the homogeneous analysis as not all components from a manufacturer are failed simultaneously. Failures are random, influenced by the failure rates. The same procedure is repeated for every column. Finally, the rows are arranged in descending order. The interpretation of Figs. 5, 6, and 7 is dealt with in Section V.

V. Results

The results are divided into three sections - homogeneous, heterogeneous, and RS analyses.

1) Homogeneous Analysis: Consider the example in Fig. 5a. The Z_C is 100% in scenario f_1 regardless of the arrangement because there are no failures. In scenarios f_2, f_3, and f_4, when aggregate manufacturers are failed incrementally for each scenario, the A_{LR} and A_{Pod} are affected badly. This is because, in the A_{LR} and A_{Pod}, all the aggregate switches from a pod are from the same manufacturer. When that manufacturer fails, the inter-rack traffic in that pod fails. In the A_{LRS} and A_{Ran}, the switches in a pod are from different manufacturers. So even when 3 manufacturers fail, there will be one functional manufacturer in the pod that ensures connectivity. The A_{Ran} may have a few pods that do not have the working manufacturer. So, a drop in the Z_C is seen. From scenarios f_5, f_6, and f_7, for the Z_C, the effect of failing core switches is not as bad as failing aggregate switches because the N_C is large. On failing both aggregate and core switches in scenarios f_8, f_9, and f_{10}, the Z_C falls. In Fig. 5b, the same trend is observed. The A_{LRS} is more predictable and offers better performance than the A_{pod} and A_{LR}. In some scenarios for the Clos-network-based topologies, the A_{Ran} outperforms the A_{LRS}. Consider the example of Figs. 8a and 8c following A_{LRS} while Figs. 8b and 8d follow A_{Ran}. The n_A and N_m are 5 and 2 respectively in both topologies. On failing the yellow manufacturer, in the A_{LRS}, alternate pods are well connected, but consecutive pods are disconnected. In the A_{Ran}, due to the absence of symmetry in the topology, connectivity may still exist. This improves the performance of the A_{Ran}. Only FT and Fb-Fab suffer from this issue. AB-FT is more robust due to its skewed arrangement. Thus, the A_{Ran} outperforms A_{LRS} only when the n_A is not a multiple of N_m.

Scenario	N_m	3TLS	FT	AB-FT	Fb-4P	Fb-Fab
n_A	-	$X > n_A$	5**	5**	4	4
Connec-tivity*	4	Good	Bad	Good	Good	Good
	5	Good	Good	Good	Bad	Bad
Max-flow*	4	Very Good	Bad	Good	Very Good	Very Good
	5	Very Good	Very Good	Very Good	Bad	Bad
* Worst case considered in this work ** n_A for medium size. Will vary for larger size.						

TABLE V: Summary of Homogeneous distribution analysis

Heat maps for all the topologies, sizes, and N_m, are compared and the results are summarized in Table V. The results here reflect the trends observed for different sizes also. 3TLS has the best performance due to higher n_A and L_{AC}. But, the cost of setting it up is high. AB-FT is superior to FT due to its inherent asymmetry in connections. But, this superiority is seen only when the n_A is a multiple of N_m. In Fb-4P, the n_A is fixed at 4 and appears as a bottleneck due to high R_{OS}. Fb-Fab in general has good performance, scalability, and redundancy.

2) Heterogeneous Analysis: Considering the example in Fig. 6, the best arrangement is A_{LRS}. From Section IV-4, to survive the current traffic, at least $66\% \times 160GB$, i.e., $104GB$, is required. This is possible only if U_{mi} is at most 30%. So, the N_m must be at least 4. To survive the traffic in the future, at least $76\% \times 160GB$, i.e., $120GB$, is required. This is possible only if U_{mi} is at most 20%. So, the N_m must be at least 5.

Comparing all the heat maps, the previous findings are confirmed. The findings of the case study for the time period 2022-2027 are summarized in Table VI. The heat maps indicate an obvious fall in performance when more components fail. Though the Z_C remains intact, the Z_F reduces quickly to values that will not sustain the traffic in the DCN. The A_{LRS} performs the best. Note that this case study is heavily reliant on accurate traffic evolution prediction.

Topo-logy	**Best-case available max-flow (Gbps)**	**Required max-flow at present (66% of best case)**	**N_m at present**	**Required max-flow in future (76% of best case)**	**N_m in future**
3TLS	320	211	3	243	4
FT	200	132	5	152	5
AB-FT	200	132	5	152	5
Fb-4P	40	26	4	30	5
Fb-Fab	160	105	5	121	6
Future = 5 years, Size = Medium, Arrangement = A_{LRS}					

TABLE VI: Case study: Required N_m as per Section IV-4

3) Robustness Surfaces: Consider the example in Fig. 7. Since this is slightly difficult to compare with other RSs, an easier plot is constructed by averaging the values along the X-axis. So, the plots for Mean-Successful Connectivity ($\overline{Z_C}$) and Mean-Average Max-Flow ($\overline{Z_F}$) are obtained as in Fig. 9. Here, the average values are plotted along the Y-axis while the percentage of failing devices are arranged in the X-axis. Generally, the variance is also plotted along with the mean. In this case, the variance carries the same results as the mean plots. Fig. 9 is for the arrangement comparison. Each curve corresponds to an arrangement. From Fig. 9a, the $\overline{Z_C}$ drops significantly at a higher percentage of failures for A_{Pod} and A_{LR}. In this example, the A_{Ran} seems to outperform the A_{LRS}.

This result is interesting and requires further investigation. Consider Fig. 10, showing the z_f distribution for the same input configuration as Fig. 7. Here, N_{m1} and N_{m2} have failed (50% failure). Even in this case, the A_{LRS} (green curve) provides a uniform available z_f of $80GB$ for each inter-rack flow in the DCN, so, Z_F is $80GB$. However, in the A_{Ran} (red curve), about 15% of the inter-rack flows have failed due to lack of connectivity, 60% of the flows have less than $40GB$ of available z_f. The rest of the 40% of the flows enjoy abundant available z_f. So, the Z_F is above $80GB$, but high priority flows may be lost due to the lack of fairness. On the other hand, the A_{LRS} guarantees fairness, predictability, and consistency. So, the A_{LRS} is preferred. Similar analyses were done for topology comparison and N_m comparison. These analyses verified the previous findings.

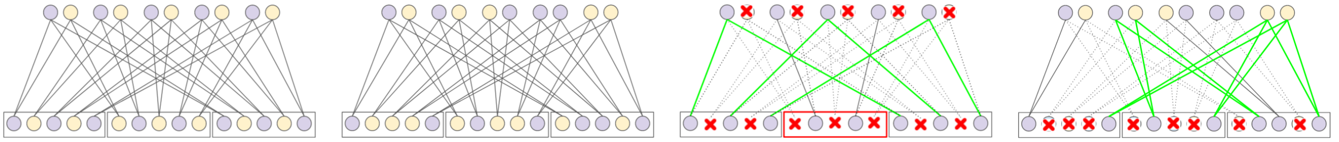
VI. Conclusion

From the results obtained, the following guidelines can be given to the DCN operators.

(G1) The number of aggregate switches per pod (n_A) must be a multiple of the number of manufacturers. Having more manufacturers than the number of aggregate switches per pod is acceptable because the probability of all but one manufacturers failing simultaneously is low.

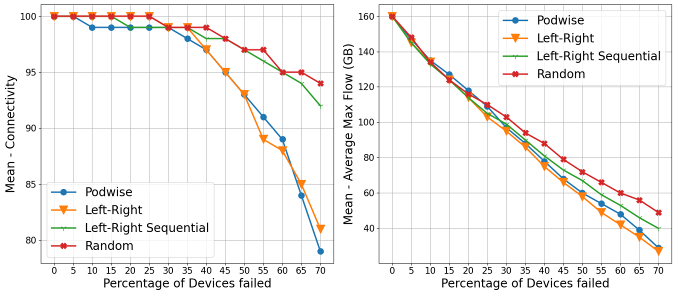
(G2) For small DCNs ($<$ 5000 servers), Leaf-Spine can be used. Cost constraint still exists. For larger DCNs, a Clos-network-based topology is needed for feasibility and scalability.

(G3) Regardless of topology, size, and arrangement the left-right sequential arrangement performs best.



(a) A_{LRS}, $N_m = 2$, No failure (b) A_{Ran}, $N_m = 2$, No failure (c) A_{LRS}, $N_m = 2$, Yellow failed (d) A_{Ran}, $N_m = 2$, Yellow failed

Fig. 8: Arrangement comparison - A_{LRS} amd A_{Ran}



(a) $\overline{Z_C}$: Arrangement comparison (b) $\overline{Z_F}$: Arrangement comparison

Fig. 9: Compressed RS - Fb-Fab, medium, $N_m = 4$

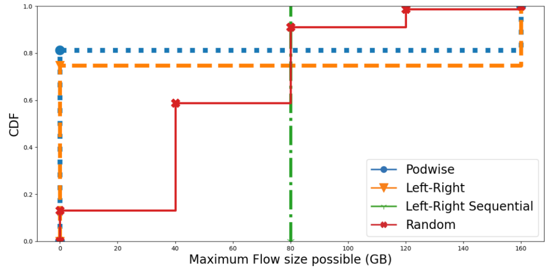


Fig. 10: z_f distribution Fb-Fab, medium, $N_m = 4$, $\{0\!:\![A,C],\ 1\!:\![A,C],\ 2\!:\![\,],\ 3\!:\![\,]\}$

(G4) The upper bound on the number of components that can be purchased from a single manufacturer ensures robustness.

(G5) Accurate traffic prediction ensures the future survival.

To study sovereignty in networks, data centers were considered an ideal first step. Different approaches to carry out the sovereignty analysis were implemented. DCNs of different sizes and topologies were modeled and studied based on the number of manufacturers used to build the DCN. The results gathered have been used to provide guidelines to the DCN operator to build a robust DCN. The same procedure can be used to link the hardware components with their software manufacturers and study the dependencies between them. In the future, these methods can be applied to larger critical networks like long-haul optical fiber networks, core networks, and later to wireless technologies.

Acknowledgment

This work has been funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under the project ”6G Future Lab Bavaria”.

References

[1] A. Weber, S. Reith, M. Kasper, D. Kuhlmann, J.-P. Seifert, and C. Krauß, “Sovereignty in information technology,” *Security, safety and fair market access by openness and control of the supply chain. Karlsruhe: KIT-ITAS. Online verfügbar unter http://www. itas. kit. edu/pub*, vol. 2018, 2018.

[2] J. Edler, K. Blind, R. Frietsch, S. Kimpeler, H. Kroll, C. Lerch, T. Reiss, F. Roth, T. Schubert, J. Schuler *et al.*, “Technology sovereignty: from demand to concept,” Perspectives-Policy Brief, Tech. Rep., 2020.

[3] P. Gill, N. Jain, and N. Nagappan, “Understanding network failures in data centers: measurement, analysis, and implications,” in *Proceedings of the ACM SIGCOMM 2011 Conference*, 2011, pp. 350–361.

[4] J. Meza, T. Xu, K. Veeraraghavan, and O. Mutlu, “A large scale study of data center network reliability,” in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 393–407.

[5] R. Potharaju and N. Jain, “When the network crumbles: An empirical study of cloud network failures and their impact on services,” in *Proceedings of the 4th annual Symposium on Cloud Computing*, 2013, pp. 1–17.

[6] R. Singh, M. Mukhtar, A. Krishna, A. Parkhi, J. Padhye, and D. Maltz, “Surviving switch failures in cloud datacenters,” *ACM SIGCOMM Computer Communication Review*, vol. 51, no. 2, pp. 2–9, 2021.

[7] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, “Inside the social network’s (datacenter) network,” in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 123–137.

[8] C. V. N. Index, “Forecast and methodology, 2015–2020,” *White paper*, pp. 1–41, 2016.

[9] R. R. Reyes and T. Bauschert, “Infrastructure cost comparison of intra-data centre network architectures,” in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2018, pp. 1–7.

[10] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, “Vl2: A scalable and flexible data center network,” in *Proceedings of the ACM SIGCOMM 2009 conference on Data communication*, 2009, pp. 51–62.

[11] N. Farrington and A. Andreyev, “Facebook’s data center network architecture,” in *2013 Optical Interconnects Conference*. Citeseer, 2013, pp. 49–50.

[12] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 267–280.

[13] Z. Hu, Y. Qiao, and J. Luo, “Atme: Accurate traffic matrix estimation in both public and private datacenter networks,” *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, pp. 60–73, 2015.

[14] T. Benson, A. Anand, A. Akella, and M. Zhang, “Understanding data center traffic characteristics,” *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 92–99, 2010.

[15] ——, “Microte: Fine grained traffic engineering for data centers,” in *Proceedings of the seventh conference on emerging networking experiments and technologies*, 2011, pp. 1–12.

[16] B. Lebiednik, A. Mangal, and N. Tiwari, “A survey and evaluation of data center network topologies,” *arXiv preprint arXiv:1605.01701*, 2016.

[17] Y. Liu, X. Wei, J. Xiao, Z. Liu, Y. Xu, and Y. Tian, “Energy consumption and emission mitigation prediction based on data center traffic and pue for global data centers,” *Global Energy Interconnection*, vol. 3, no. 3, pp. 272–282, 2020.

[18] M. Manzano, F. Sahneh, C. Scoglio, E. Calle, and J. L. Marzo, “Robustness surfaces of complex networks,” *Scientific reports*, vol. 4, no. 1, pp. 1–6, 2014.