Cost Optimization by Energy Aware Workload Placement for the Edge Cloud Continuum*

Rickard Brannvall[1,2], Tina Stark[1], Jonas Gustafsson[1], Mats Eriksson[3], and Jon Summers[1]

[1]RISE Research Institutes of Sweden AB
[2]Luleå University of Technology
[3]Arctos Labs Scandinvia AB

October 2022

*This work was supported through the Datacenter Innovation Region (DIR) project at Luleå University of Technology

Abstract

This report investigates the problem of where to place computation workload in an edge-cloud network topology considering the trade-off between the location specific cost of computation and data communication.

1 Introduction

Recently, edge computing has been proposed to support new digital services such as live video processing, 5G, artificial intelligence, virtual reality, augmented reality, robotics, IoT, additive/incremental manufacturing, and direct monitoring and control of cyber-physical systems. This is based on the assumption that these applications place higher requirements on the digital infrastructure in terms of latency demands, privacy restrictions, as well as increased capacity for high-performance computation and data communication. Edge computing will also reduce the requirement for data transmission infrastructure.

The provisioning of such edge capacity introduces cost considerations that need to be taken into account in the design of a distributed edge-cloud infrastructure comprised of a network of edge data centers combined with traditional centralized cloud compute. The cost of computation can vary between different locations in such a network topology, for example, because of economies of scale aspects that may make smaller data centers less efficient [1, 2]. It is well known that the worldwide electricity consumption of data centers is significant, which is why efficiency matters so much, especially when considering the carbon footprint that it implies. Data communication further increases the energy consumption to a degree comparable to that of the computation itself [3].

1.1 Edge or cloud?

The concept and terminology of edge have been discussed, and confusion may arise on the term of edge compute. In this report, the edge is considered to be small data centers furthest out in the network, close to the end-user devices.

When deciding the placement of workloads, there are more or less two reasons, either due to constraints or optimization. One example of constraints is when the application is latency critical and will not fulfill its purpose if it is located too far away from the device. Examples here are machine control in the manufacturing industry and traffic support services. Another example of a constraint is privacy/confidentiality considerations that mandate the processing of data within some physical boundaries. In the absence of such hard constraints, it may sometimes be more optimal to select a certain location over another. Therefore, this report aims to investigate the question:

Where is the most cost-efficient location to place a specific workload at a particular time in a dynamic edge-to-cloud continuum?

Cost efficiency is here intended to be read in the wider meaning that factors in energy efficiency and environmental footprint. It relates to the two aspects of 1) the investment decision and design of an edge data center infrastructure and 2) the optimal placement of a task such that the computational assets are best utilized at any given time. The first aspect includes the overall cost of building and operating such infrastructure, while the second relates to the marginal cost of producing an additional service in a given edge-cloud infrastructure.

This report explores two models for the cost savings that can be made by moving data and computational tasks to different locations in the edge-to-cloud continuum, both up/down between edge and cloud but also sideways between different edge data center nodes. First, we re-examine a cost model that was earlier presented by one of the authors [4] to elicit under what conditions the savings on data communication with the cloud can justify the use of a potentially less efficient edge data center close to the user. The second model assumes an operational sweet spot for (edge) data centers and investigates when it is efficient to redistribute computational load to the edge to bring its hardware closer to such an operational sweet spot.

Limitations. This report analyses the marginal costs for additional workloads and does not intend to capture a full life cycle cost analysis. An optimization perspective on edge compute is considered, assuming that premises are already built for other reasons, such as the deployment of 5G, with edge data center hardware provisioned to meet requirements imposed by aforementioned constraints (e.g., it is designed to handle worst-case volumes of latency-critical workloads that must be executed locally at the edge).

We then assume that on top of these, there are additional workloads that originate in the same locales but that can be deployed with perfect discretion either at the edge or at a centralized cloud data center.

1.2 Cost of computation

Computation costs energy, and a common measure of energy efficiency is Power Usage Effectiveness (PUE). PUE is a ratio of total energy consumed in the data center over the energy consumed by the IT hardware, where the total energy in the nominator includes IT, cooling, power transformation, and all other sources associated with the facility[5]. In table 1, the Power usage effectiveness (PUE) values from two studies are presented. The PUE in [6] from 2014 is a projection of PUE numbers in 2020. In [7] from 2017 with actual numbers, it shows an almost flat curve for different IT-rated loads. This suggests that the IT capacity has no actual impact on the efficiency of the facility and that the PUE could be assumed to be more or less the same for data centers of different sizes. According to [8], the average PUE in the world for 2021 is 1.57.

Table 1: Data center types and the number of servers used.

Slogan	No of servers	PUE [6]	PUE [7]
Regular	15,000	1.57	1.64
Small	1,500	1.85	1.87
Micro	15	2.27	1.73

As mentioned in [7], data centers that are located in the Nordic countries show to have a lower PUE due to the colder ambient conditions where free cooling can be used more. For example, the Hydro66 data center located in

Boden, northern Sweden, states a PUE of 1.07 [5]. Experimental data center designs show even lower numbers, less than 1.04 [9]. And in another research project at RISE, where tests on an experimental edge module was done, a PUE of 1.09 was obtained.

1.3 Cost of communication

The cost of communication, that is, to move data from the device to the data center, is highly discussed. Aslan et al. [10] has done an investigation on different estimates of data communication costs from 2000 to 2015, presented in Figure 1, where the numbers decrease for each year in an exponential fashion. As a similar (exponential) trend is reported for the cost of computation [3], one can conclude that a comparison of the two costs is relatively stable across time. If the case were otherwise, over time, one would eventually end up dominating the other, making any optimization based on their trade-off trivial.

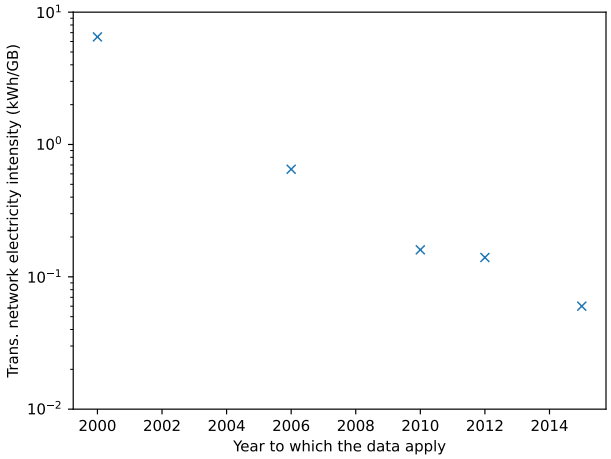

Figure 1: Data communication electricity intensity from [10].

Malmodin et al. 2012 [11] estimates the average energy consumption to 0.08 kWh/GB while a few years later in Malmodin 2020 [12], the marginal energy cost is estimated at a much smaller value of 6.7e-5 kWh/GB.

We understand that the discrepancy between the two estimates reported in [11] and [12] can be explained by different assumptions that each expresses an alternative perspective on the communication cost: the former (and higher) estimate is calculated as an average of the total cost including the communication infrastructure, and therefore takes into account the cost of a potential expansion of the infrastructure that a substantial increase in total communication might warrant if the limit of data transfer capacity is met. The latter (and much

lower) estimate represents a marginal cost of communication given a fixed data communication infrastructure.

2 Modelling results

2.1 Model 1: Compute vs. Communication

This model is based on work by one of the authors [4] that was earlier presented at The Edge Event 2020. In this section, it is developed further with an updated analysis of the cost of data communications.

Review of cost modelling

Server model. The marginal cost of compute has been analyzed in [13] by using Standard Performance Evaluation Corporation server efficiency rating tool benchmarks [14]. It shows that it is still around 25% power consumed at 10% workload. From that point, the consumption for different workloads changes depending on the benchmarks within the tool. Eriksson (2020) [4] calculated an average over the various benchmarks of the tool, arriving at a profile that is nearly linear with a maximum power at around 450 W. If we simplify the utilization curve to a linear dependency from P_{idle} to P_{max}, the power q consumed by a single server at a certain utilization u can be expressed as

$$q = a + b\,u = P_{idle} + (P_{max} - P_{idle})\,u \tag{1}$$

This curve can be seen in figure 2 and is assumed to have a maximum power $P_{max} = 450$ W. Using the rule of thumb from [14] that around 25% power consumed at 10% workload we find $a = P_{max} = 75$ W, with the slope $b = 375$ W.

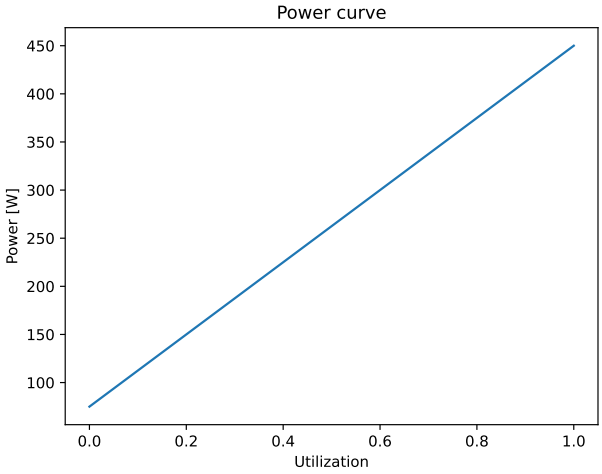

Figure 2: Power curve

Workload definition. The server assumed for this exercise is running equipped with 16 cores and can then produce 16 CPU-h per hour of operation. The definition of a workload then becomes a thread that occupies one full CPU core for a given unit of time. Adding an extra workload (thread) to a server then corresponds to an increase in its utilization by 6.25% and, equivalently, an increase of 23.4 W in terms of power.

Marginal cost of compute. As described in [6], the PUE may vary between data center sizes, and an increase in server power needs therefore to be multiplied with its corresponding PUE. The marginal cost imposed by an additional CPU-h can then be calculated by

$$[\text{CPUh cost}] = \text{PUE}\,\frac{(P_{max} - P_{idle})}{n_{\text{CPU}}}\,\frac{[\text{energy price}]}{1000} \tag{2}$$

where n_{CPU} is the total number of CPUh per hour delivered by the server, and energy price is the cost of electricity which is assumed to be 0.1$/kWh for our calculations. (The constant is just to bring the cost into units of kWh.) This translates into a dollar cost per CPUh of 2.34e-3 at a minimal PUE of 1.0, but of course, has to be scaled proportionally for any real PUE > 1.0.

Marginal cost of communication. Recall from section 1.3 that we have two views of the cost of data communication. The marginal cost when transporting an additional workload on an existing network infrastructure estimated[12] at 6.7e-5 kWh/GB is much lower than an averaging estimate[11] of 0.08 kWh/GB based on distributing the entire cost of that infrastructure on all traffic equally.

Assuming the same cost of electricity, these cost becomes 6.7e-6 $/GB and 8e-3 $/GB, respectively.

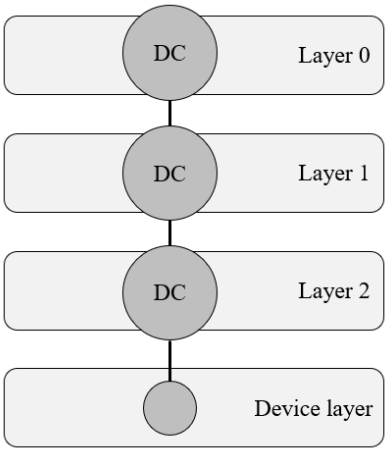

Figure 3: Topology

Network topology. The topology analyzed for model 1 can be seen in Figure 3 which is composed of three hierarchical layers, where each layer consists of multiple data centers. In this simplified topology, we allow connections between adjacent layers for but ignore interconnections within layers or that skip layers to obtain a tree structure.

The sizes of the data centers present at each layer are such that layer 0 corresponds to a central data center of regular size as defined in Table 1. While the data centers at layer 0 may deliver cloud services supported by large economies of scale, layer 2 instead is located closest to the end-user device and may correspond to micro-sized (Table 1) edge data centers delivering compute at the lowest latency. Layer 1, made up of small-sized data centers (Table 1), is somewhere in between the cloud and edge.

Workload definitions

The workloads can be classified based on their requirements for compute power and data transport, as is done in Table 2 that we borrowed from [4].

The three workloads that are listed in Table 3 and further described in the paragraphs below have been selected from each of the three interesting categories that are highlighted in Table 2 to provide examples of the varying effect of placement in the simulation exercise for model 1 that are presented later in this section.

Web Server. Based on Apache bench, the amount of data transferred has been captured for the rendering of an example web page as well as the compute

Table 2: Different types of workloads (from [4])

	Low amount of compute	**High amount of compute**
Low amount of data	Low cost so placement is less sensitive	Compute cost dominate central location
High amount of data	Data transport cost dominate, placement –closer to edge	Need for more detailed modelling

consumed for doing so. The relation between data and compute is the interesting part to analyze, and where in the topology that work should be best placed. In this case, it is also clear that data for the web page needs to be available. This aspect has been omitted in the analysis, assuming that all data needed are available at all locations and no additional transfer need to take place. This workload is quite data intense in relation to the compute work that is required.

Video Compression. Here a benchmark from SPEC [14] was used that compresses a YUV file into a MPEG-4 file. In this analysis, we assume that the YUV file emerges from the device and is sent upstream for compression. The compressed result is then sent further upstream to Layer 0. So, there is a data transport input, a compute workload, and a data transport output to take into consideration. This workload is both data and compute-hungry.

Image Manipulation. This is another workload in the SPEC benchmark series. It operates on a 2068x1380 pixel image as input and manipulates the image in a series of operations, resulting in a 3299x5002 pixel image that is assumed to be sent upstream towards layer 0. This workload is very compute-intensive at a moderate need for data transport.

Workload summary. Table 3 gives a summary of the different workloads, where (in) and (out) in the table refers to data that is the input or output of the computation. Note that it is the ratio between compute and data transport that is interesting for this exercise. However, the needed compute and data relate to very different work and should not be compared to each other in absolute terms.

Table 3: Summary of the workloads used.

Workload	CPU/h	GB in	GB out
Web server	0.001	0.0059	0.000
Video compression	0.430	0.0700	0.009
Image manipulation	1.350	0.0090	0.050

Comparison of low and high communication cost

The model was run for the three different workloads with both the low and high communication costs, mentioned in section 1.3, and a fixed PUE of 1.57. Figure 4 show the savings per workload of putting the work on layer 1 or 2 compared to layer 0. Here, the higher communication cost is considered. For the Video compression and Web server workload, the results show that the savings are up to 65% and 32%, respectively, at layer 2, the edge. For the Image manipulation workload, the result is the opposite, with a negative number for layers 1 and 2.

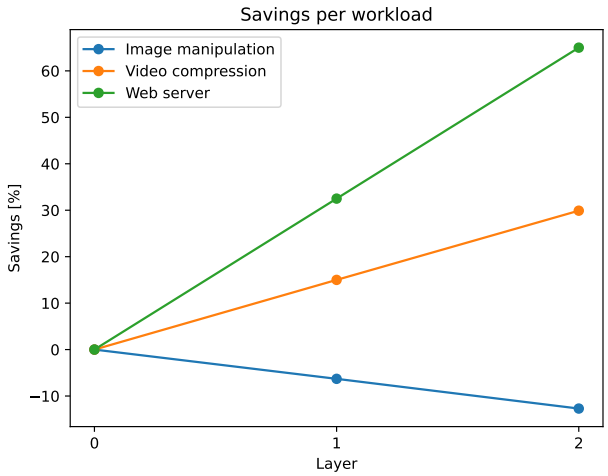

Figure 4: Savings per workload with the (higher) communication cost corresponding to the average case from [11].

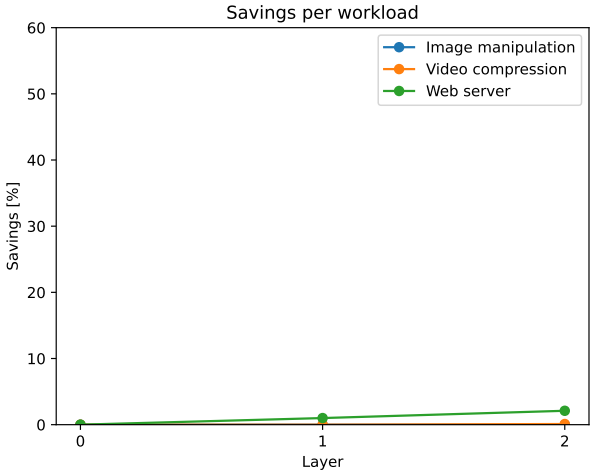

Figure 5: Savings per workload with the (lower) communication cost corresponding to the marginal case from [12].

Figure 5 shows the same calculation but with the lower communication cost. Here, the only workload that has a saving is Video compression at layer 2, although it is very small at 2.1 % and should thus not be considered significant. For the other workloads, the cost is more or less the same for each layer, reflecting the assumption of a data communication cost that is almost negligible.

Comparison of flat or skew PUE profiles

The model was run for the three different workloads on each layer first with a fixed PUE of 1.57, and then with a variable PUE that depends on the assumed data center size at each layers from table 1. For these calculations the (higher) communication cost of 0.08 kWh/GB based on averages [11] was used.

The results for the variable PUE case are displayed in Figure 6, exhibiting the same trend as we see in Figure 4 for the fixed PUE case (assumed in the previous exercise for high communication cost calculation).

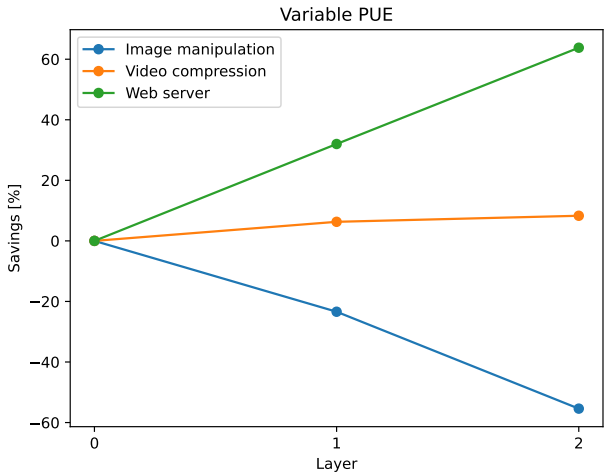

Figure 6: Savings per workload for the case of variable PUE.

The workloads Video compression and Web server shows savings at layer 1 and 2 whereas Image manipulation should be kept at layer 0. Looking at Video compression, the fixed PUE calculations show a higher saving at layer 2, at 65% compared to 10% for a variable PUE. The Web server workload has more or less the same saving at layer two for fixed and variable PUE.

Sensitivity analysis. For this we refer to the sensitivity analysis for the model presented in Eriksson [4] used as a basis for Model 1 of this work.

2.2 Model 2: Effect of placement

This section focuses on the efficient placement of computational loads at either a cloud or edge data center, which we for this exercise assume to have the same instantaneous PUE curve. If the cost of moving data is very small, why does it matter where it is executed? A first guess may be that we save more by placing the workload in the cloud which is optimized to run at an operational sweet spot almost regardless of the total utilization (thanks to its superior economies of scale). Or is it better instead to opportunely place additional workloads at the edge data center with the objective to bring its hardware closer to its sweet spot. Furthermore, can we save on hardware in the cloud by making better use of the edge data center capacity?

Power utilisation of the edge data center
The model in this section takes as input a power utilization curve for the entire data center, which includes both power from the IT equipment and from

supporting systems such as heat removal. The relation between the (computational) utilization of the data center and the power it consumes is of course impossible to capture in a one-variable curve, as it may have complex time-dependencies caused by e.g. thermal inertia or external driving factors such as ambient temperature. We will however make use of a simple model obtained from the linear model of Figure 2 for IT power utilization when combined with a cubic law for the power required for heat removal. Details of the derivation are given in Appendix A – the resulting curve is displayed in Figure 7. We have assumed that our edge data center consists of 24 computer servers – that is, its size is about half a rack.

Note that by taking the ratio of the curve in Figure 7 by the IT power curve of Figure 2 we obtain PUE (as a function of utilization). We can thus calibrate the curve to fall within a reasonable range of PUE values.

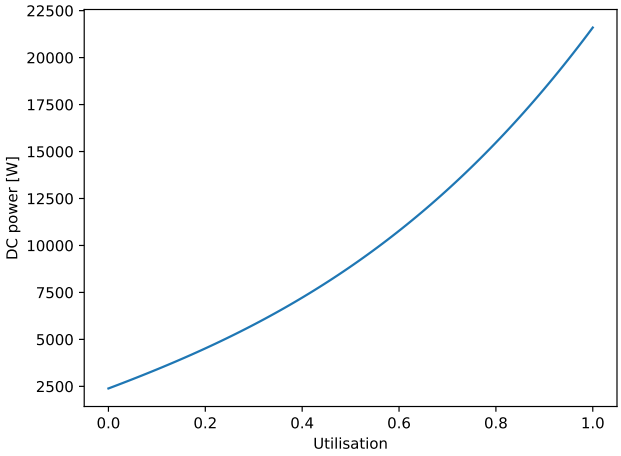

Figure 7: The total power utilization curve for the edge data center follows a cubic law in the stylized example we explore in our simulations, where utilization between 0 and 1 of the computational resources is used.

Workload thread. For the first simulation exercise in this section, we further consider the cost of communication to be zero. This means that the cost of running the workload is entirely dependent on the actual efficiency of compute. From the total power-utilization curve, we can derive the average power-utilization curve $p(u)/u$. For this study we assume that each task occupies 100% of the capacity of a computational core of a multiple CPU core server – we call this a thread. A thread then becomes our smallest unit of increment for utilization, which for our example edge micro data center comes in at approximately 0.26% since we assume 24 servers, each having 16 cores.

Figure 8 shows the cost of a single task (that each fully occupies a thread)

at different utilization levels. We note that for our parameter choices, this curve has a minimum around 60% utilization – we call this the operational sweet spot and note that here each task cost is approximately 37.5 W.

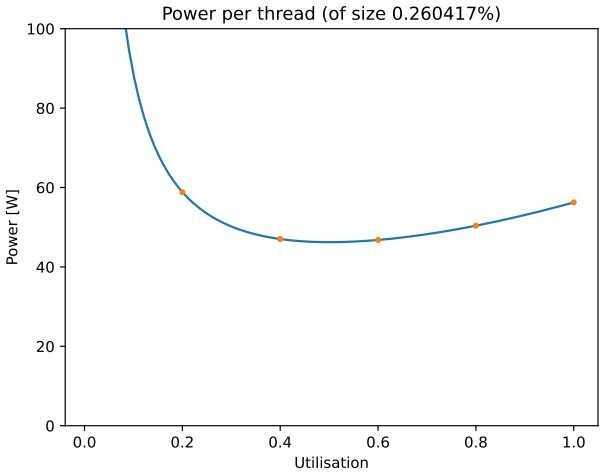

Figure 8: Average power consumed per thread for a hypothetical edge data center (solid line). The orange dots indicate workload levels in steps of 20% utilization.

Load placement flexibility. We here assume that computation workloads are generated in a local area where the edge data center is placed. Some of these tasks MUST be placed at the edge data center, for example, because they require low latency or process confidential information that can not be entrusted to the cloud. Other workloads, although generated in the vicinity of the edge data center, can as well be run in the cloud. We, therefore, assume two categories of workloads: *fast* and *flex*. It is to serve the former that the edge data center exists. For the latter category, however, we want to explore placement strategies that are optimal for energy efficiency and other sustainability objectives.

Single edge data center. A *flex* type of workload should be placed at edge or cloud depending on how it affects the total energy cost, which in turn depends on how many other workloads that are active at the same time. Table 4 shows the total cost for different combinations of workload placement between edge and cloud, such that rows indicate the resulting utilization at edge and columns show the workload placement in the cloud (in units normalized to the edge data center capacity). We can read the different workload placement alternatives for a fixed total workload on the anti-diagonals of the table (since for these, the sum of the row index and column index is constant). For simplicity, the table is given in increments of 20% of the edge data center's capacity - smaller

Table 4: Total cost in Watts of different workload placement combinations: rows indicate the edge data center utilization and columns the amount of workload placed on the cloud (in units proportional to the edge data center size).

edge / cloud	0.0	0.2	0.4	0.6	0.8	1.0
0.0	1823	4703	7583	10463	13343	16223
0.2	3780	6660	9540	12420	15300	18180
0.4	6008	8888	11768	14648	17528	20408
0.6	8640	11520	14400	17280	20160	23040
0.8	11813	14693	17573	20453	23333	26213
1.0	15660	18540	21420	24300	27180	30060

increments are of course possible for a more granular table. Note that for our parameter choices, it is optimal to fill the edge data center up to the sweet spot of 60% utilization, after which it is more economical to place the workload in the cloud.

Another view on the impact of placement for a single edge data center is provided in Table 5. Each row assumes a fixed utilization of the edge data center and then looks at the additional cost of a *flex* workload that can be placed either at the edge or in the cloud. The incremental cost for each choice is then shown in the two columns of the table. Also here we see that we should only place *flex* workload in the cloud after we have filled the edge data center up to its sweet spot, which for our parameter choices is at 60% utilization.

Table 5: The average cost in Watts of different workload placement strategies where the rows indicate the present edge data utilization level.

edge utilisation	place in cloud	place at edge
0.0	37.5	24.3
0.2	37.5	27.0
0.4	37.5	31.4
0.6	37.5	37.5
0.8	37.5	45.5

Time-varying workload pattern. We assume that the edge data centers are appropriately designed and distributed, such that the amount of *fast* workload at any time never exceeds the capacity of the edge data center. To justify this assumption we argue that otherwise, if requests for larger workloads were to occur in a local area, it would simply have to be split up and more edge data centers be placed there to meet the demand. We also, rather arbitrarily determine that the occurrence of *flex* workload that can be generated in an area is to be the same amount as for *fast* workloads.

The amount of workload of each type that is generated for a locale then

follows a certain probability distribution, where marginally we use the beta distribution such that for *fast* workloads we have

$$u^{fast} \sim \text{beta}\left(\alpha^{fast}, \beta^{fast}\right)$$

where α and β are the parameters of the beta distribution in its standard formulation (see e.g. Wikipedia [15]). Correspondingly, for *flex* workloads we have

$$u^{flex} \sim \text{beta}\left(\alpha^{flex}, \beta^{flex}\right)$$

The baseline distribution that we use for this study will be a left skew distribution with parameters $\alpha = 1.5$ and $\beta = 3.0$. Alternatively, for the sensitivity analysis below, we will use the uniform distribution which is obtained by setting parameters $\alpha = 1.0$ and $\beta = 1.0$. For simplicity, we choose the same distribution for *flex* and *fast* workloads in all the simulations that follow below.

Single cycle assumptions. We assume that workloads do not persist over multiple cycles. That is, a workload thread initiated at time t will complete before the next time period leaving both the edge data center and the cloud free to take on new workload. Although allowing for persistent workloads would make the simulation more realistic, it would complicate the optimization that is used for workload placement without increasing the model's capacity to examine the central problem of this study.

Homogeneity assumptions. For this study we assume that the workload distribution is homogeneous across time and locales, such that we use the same distribution for each workload type regardless of the time t and locale index j. This assumption is examined in Appendix C.

Correlation assumption. To simulate bursts of activity that affects both *fast* workloads and *flex* workloads we allow dependence when we jointly draw u^{fast} and u^{flex}. We let the correlation between flex and edge random draws ρ be modeled by a two-variate Gaussian copula [16]. The mathematical details of this are left for Appendix B, but we note here that the copula formulation concerns the joint distribution of these two variables and does not change the marginal distributions given earlier in this section. The dependence between the variables is controlled by the (linear) correlation parameter of the Gaussian copula. We set this parameter $\rho = 0.7$ as default for our simulations. This assumption is examined in Appendix C.

Workload placement strategies. Here we investigate three modes of operation for workload assignment according to:

default: all flex workloads are placed on the cloud

local: flex workloads are placed on the local node or shipped to the cloud

global: any flex workload can be placed on any local node, or cloud

and for each, workloads are placed opportunistically such that energy cost is minimized. Note however that the *default* mode offers no opportunity for optimization as *flex* workloads are always placed on the cloud (which we assume has infinite capacity for absorbing workload). The difference between the other two modes are that in *local* mode each edge node is filled with workloads until it is more beneficial to move its excess workload to the cloud, while for *global* mode we allow *flex* workloads that originate in one local node to be moved opportunistically to any other node that has free capacity as long as it reduced the total energy cost.

Simulation of a fleet of edge data centers. Each time period t we draw the workload requirements

$$(u_{t,j}^{fast}, u_{t,j}^{flex}) \sim \Pi(\alpha, \beta, \rho)$$

for all locales j under the assumption of having homogeneous distributions across time and locales, where Π is the joint distribution capturing the dependence between the two workload categories.

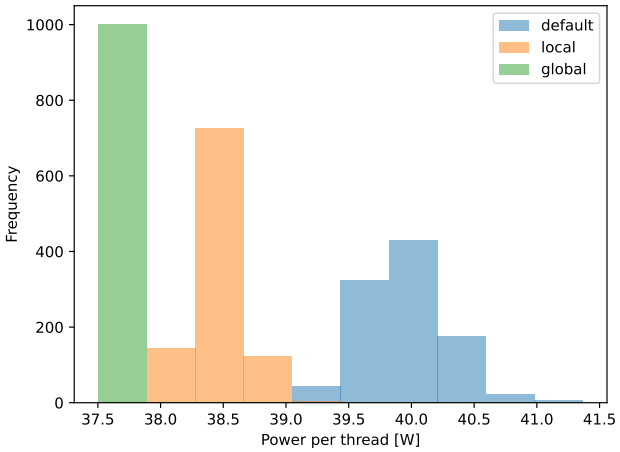

Figure 9: Histogram of the per thread energy cost under the three different workload assignments modes. The two strategies that place more flex workload at the edge reduce average cost from 40 W to 38.5 W and 37.5 W, respectively.

Figure 9 show results from a simulation for a fleet of $N = 100$ identical edge data centers for $T = 1000$ time-steps under the simple cubic power-utilisation law and other assumptions discussed above (left skew distribution and positive correlation). It is presented as a histogram showing frequency on the y-axis and

the per thread energy cost under the three different workload assignments modes on the x-axis. Note the separation between the three modes, with a clear trend of lower power requirements per thread for the more advanced optimization strategies. On average the *local* mode and *global* mode shows savings of 3.7% and 6.0% compared to *default*, respectively.

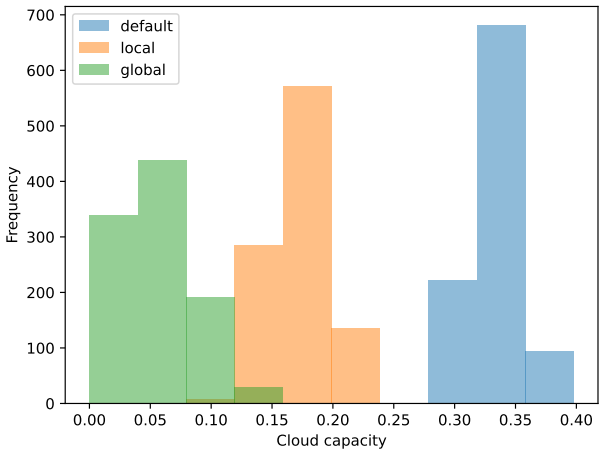

Figure 10: Histogram of the capacity claimed from the cloud for workloads that are generated locally at the edge and then moved to the cloud. By using the edge data center more optimally (*local*, *global*) the claim on the cloud is reduced.

Another view on the same simulation is provided by Figure 10, which shows the capacity claimed from the cloud in units of the total edge capacity, that is, the amount of workload that are generated locally that are moved to run in the cloud. The histogram distributions show some overlap, but it is clear that the two alternative workload placement strategies shows advantage over *default* since the claim on the cloud is significantly lower. This means that workloads can be run efficiently at the edge to free up capacity at the cloud. In our simulation example with $N = 100$ edge data centers we see that the claim on the cloud can drop from almost 35% down as low as 5–15% (measured in terms of total edge data center capacity). Scaled to our example of 100 edge data center nodes each with 24 servers, this translates to a saving in hardware of up to 700 servers.

Compared to the default strategy, the opportunity provided by the *local* mode and *global* mode for saving on hardware is on average 48.2% and 83.6% respectively, since over time one would have to invest less in the cloud.

2.3 Combining the models

In the basic configuration of the model in the previous section (model 2) we assumed that the cost of moving workload is zero, that is, we assumed that all input and output communication associated with the computation takes place free of charge. Section 1.3 contains a more detailed discussion of the cost of moving data. For this section we take the communication cost of the least data hungry applications explored in Model 1 – Image manipulation – as we already saw a benefit of moving the other task to the edge. To place the communication cost on the same footing as the other cost of model 2, we translated into a power requirement of approximately 5 W, by calculating the ratio of communication cost to computation cost given by Table 3.

The communication cost is paid once for moving a workload from one place to another in the network topology (regardless of distance in terms of layers). Note that for this model we consider a two layer topology consisting only of cloud and edge, and hence we ignore the intermediate layer of Model 1, and furthermore allow full interconnectedness between nodes in the edge layer. We have then that when a workload that is generated near an edge node is moved to the cloud it takes a 5 W cost. A workload that is moved from one edge node to another edge node also take the same cost of 5 W.

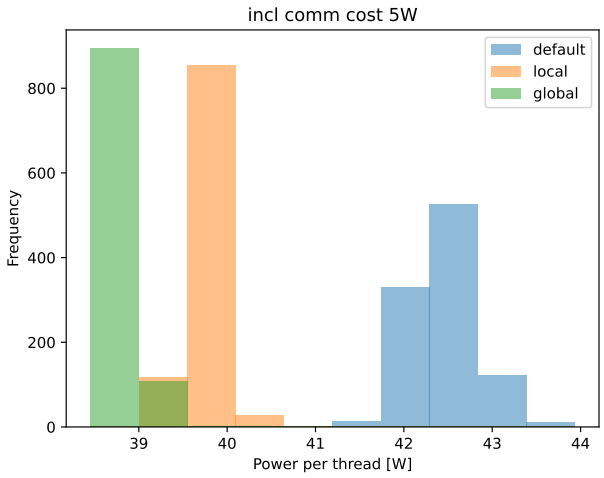

Figure 11: Results from combining Model 1 and Model 2 such that a communication cost of 5 W is counted for moves of any distance in the network topology.

Effect of communication cost. Figure 11 shows the power savings for following the same optimization strategies as for model 2. The benefit of using the edge also for fast workloads (*local* and *global* modes) is increased to 6.4% and 8.5% , respectively, compare to the *default* strategy. We also note that the relative benefit of *global* over *local* optimization remains – there is no in-

creased cost of moving workloads between edge data centers. Other assumptions about how to account for distance in the network topology may lead to different conclusions, why we want to be careful with any hard conclusions from these result.

The capacity claim on the cloud is not affected by the communication cost in this simulation (which is why we do not present any figur for this).

3 Discussion

The exercises with Model 1 in Section 2.1 outlined the benefit of placing data hungry workloads at the edge closer to the end user devices, as exemplified by the Web Server application that showed a saving above 60% when we assumed the communication cost that averages over all traffic. For the other workloads that assumed a moderate or low data communication component, the savings were no longer material (Video Compression application) or even reversed into a significant extra cost (Image Manipulation application). These results were quite robust to the assumptions of PUE, showing similar savings (or losses) both for PUEs that were fixed and that varied accross the different data center sizes. The effects however almost entirely disappear when we use the other (and much lower) communication cost estimate, based on marginal cost of adding an additional data packet on a communication network that is already up and running. A limitation of Model 1 is that it only considers the marginal cost of computation – the cost keeping servers on in a ready state is ignored.

Model 2 in Section 2.2 considers the average cost of computation at an edge data center operating at different utilization levels. Simulation results indicate that around 4–6% of the energy cost can be saved by placing workloads at the edge if this is done to bring it closer to its operational sweet spot. Furthermore, placing workloads at the edge reduced the requirements on the cloud such that savings of up to 50% of cloud hardware could be obtained according to the model estimates, pointing in the future direction of a leaner cloud.

OPEX and CAPEX. Note that while the energy saving are modest, the simulation results from model 2 indicate that opportunities for savings on hardware can be more substantial. As the former saving corresponds to operational expenses and the latter to investment in equipment, one could draw the conclusion that the optimisation does more for CAPEX than for OPEX. To be comparable the savings should be expressed in (absolute) dollar terms. This is however beyond the scope of this report, as it would require further assumptions on the price of electricity, hardware procurement costs and amortisation models. Furthermore, for a detailed analysis one may also be required to allow the cost of operation to vary between different locales to reflect regional heterogeneity in the price of electricity and/or environmental circumstances that affect PUE.

LCA implications. If also LCA costs of the equipment is accounted for the benefits of utilising edge equipment more optimally should be even more ap-

parent. A complete analysis would translate all energy and LCA costs into an equivalent measure (for example CO_2 equivalents) so that pros and cons of each strategy could be evaluated holistically. However, such comparison will depend very much on details of the modelling, such as the energy mix for the regions were the hardware is produced and the region were the equipment is operated, respectively. We therefore leave such analysis for future work.

4 Conclusions.

The results from the exercises with the two models point to the conclusion that distributing the workloads over edge data center nodes is motivated by sustainability considerations – we saw savings both in terms of electrical energy and hardware equipment – that would translate into reduced GHG emissions and a lower life-cycle footprint.

We recall that a limitation of this study is the assumption that workloads can be moved freely around in the cloud-edge continuum with perfect discretion, which is of course only possible in certain business models for the underlying digital services and ownership of hardware resources. If market solutions do not emerge naturally, one may consider how policy and other incentives can be changed to support this development. How can we stimulate energy and resource optimality in society, at regional as well as global scales?

References

[1] United states data center energy usage report. Technical report, 2016.

[2] Rolf Harms and Michael Yamartino. The economics of the cloud. Technical report. Microsoft.

[3] George Kamiya et al. Data centres and data transmission networks. Technical report, 2022. URL `https://www.iea.org/reports/data-centres-and-data-transmission-networks`. International Energy Agency, Paris.

[4] Mats Eriksson. Cost modelling of edge compute. Presented at The Edge Event 2020 (non peer reviewed), 2020.

[5] Hydro66. What is pue? URL `https://www.hydro66.com/colocation-services/what-is-PUE/`.

[6] Consumer Research Associates. Energy efficiency policy options for australian and new zealand data centres. Technical report, 2014.

[7] P Bertoldi, M Avgerinou, and L Castellazzi. Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency. Technical report, Luxembourg, 2017. URL `https://publications.jrc.ec.europa.eu/repository/handle/JRC108354`.

[8] D Bizo, R Ascierto, A Lawrence, and J Davis. Uptime institute global data center survey 2021. Technical report, 2021. URL `https://uptimeinstitute.com/resources/asset/2021-data-center-industry-survey`. Uptime Institute.

[9] Sebastian Fredriksson, Jonas Gustafsson, Daniel Olsson, Jeffrey Sarkinen, Alan Beresford, Matthew Käufeler, Tor Björn Minde, and Jon Summers. Integrated thermal management of a 150kw pilot open compute project style data center. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 1443–1450, 2019. doi: 10.1109/INDIN41052.2019.8972145.

[10] Joshua Aslan, Kieren Mayers, Jonathan Koomey, and Chris France. Electricity intensity of internet data transmission: Untangling the estimates: Electricity intensity of data transmission. *Journal of Industrial Ecology*, 22, 2017. doi: 10.1111/jiec.12630.

[11] Jens Malmodin, Dag Lundén, Mikael Nilsson, and Greger Andersson. Lca of data transmission and ip core networks. In *2012 Electronics Goes Green 2012+*, pages 1–6, 2012.

[12] Jens Malmodin. The power consumption of mobile and fixed network data services. In *Proceedings of Electronic Goes Green 2020*, pages 87–96. Fraunhofer Verlag, 2020.

[13] Norbert Schmitt, Jóakim von Kistowski, and Samuel Kounev. Emulating the power consumption behavior of server workloads using cpu performance counters. pages 157–163, 2017. doi: 10.1109/MASCOTS.2017.17.

[14] Standard Performance Evaluation Corporation. Spec. URL `https://spec.org/`.

[15] Wikipedia. Beta distribution — Wikipedia, the free encyclopedia, 2022. URL `https://en.wikipedia.org/wiki/Beta_distribution`.

[16] Wikipedia. Copula (probability theory): Gaussian copula — Wikipedia, the free encyclopedia, 2022. URL `https://en.wikipedia.org/wiki/Copula_(probability_theory)#Gaussian_copula`.

[17] Abe Sklar. Fonctions de repartition a n dimensions et leurs marges. pages 229–231. Publ. Inst. Statist. Univ. Paris 8, 1957.

A Power Utilization

We assume that the power consumption from IT, P_{IT}, follows a linear curve (as commented on in Section 2.1),

$$q(u) = a + bu,$$

and further assume that the power required for heat removal, P_{HR}, scales with the cubic fan law, that is,

$$p(q) = c + dq^3,$$

in a range between the minimum and maximum heat loads experienced by the system. This can be considered a reasonable approximation for the components of a data centers heat removal system that are concerned with moving air or liquid, such as fans in a [HVAC] or pumps in a [chiller] loop.

Total Power. We can now derive for the total power consumption, P_{TOT},

$$\begin{aligned} P_{TOT} &= P_{IT} + P_{HR} \\ &= q + p(q) \\ &= a + c + bu + d(a + bu)^3 \end{aligned}$$

which again is a cubic law.

Calibration to PUE. We can similarly write for the Power Utilization Effectiveness (PUE) that

$$\text{PUE} = \frac{P_{TOT}}{P_{IT}} = 1 + \frac{p(q)}{q}$$

and calibrate parameters such that the implied PUE of the system stays above 1 and below an assumed maximum, which is 1.5 for our example. Given the model for IT power consumption we find that the following parameters give a reasonable calibration for our purpose,

$$\begin{aligned} c &= 0.05, \\ d &= 0.45/q_{MAX}^2, \end{aligned}$$

with the implied power utilization curve displayed in Figure 7.

B Copula function

Sklar's theorem states that any multivariate joint distribution can be simplified and written in terms of univariate marginal distribution functions together with a copula that describes the dependence structure between the variables[17]. Copulas are popular in high-dimensional statistical applications since they allow

the separate modelling and estimation of the marginal distribution functions and the dependence structure of a random vector. The popular Gaussian copula is a distribution over the unit hypercube $[0,1]^d$, which is constructed by transforming random draws from a multivariate normal distribution over $\mathbb{R}^d$

$$\vec{x} \sim N(\Sigma, \vec{\mu})$$

where $N(\Sigma, \vec{\mu})$ is the joint distribution function of a multivariate normal distribution with mean vector zero, $\vec{\mu} = \vec{0}$, and covariance matrix equal to the correlation matrix Σ. Individual entries of X are then transformed to the unit interval by the cumulative distribution function of a standard normal,

$$v_i = \phi(x_i),$$

to obtain $\vec{v}$, a vector in the unit cube. As each entry v_i of the vector is uniform distributed, we can further transform these to the marginal distribution of our choice by the application of its inverse cumulative distribution function, for example for the Beta distribution we have

$$u_i = F^{-1}_{\alpha,\beta}(v_i),$$

with parameters α and β that control the shape of the probability distribution. This is also the the choice of marginal distribution function in the main text, which together with the copula described in this section defines the joint distribution $\Pi(\alpha, \beta, \rho)$ with correlation parameters ρ for the dependence between fast and flex loads at a edge data center node.

C Model 2 Sensitivity Analysis

Effect of linear power-utilization curve. Figure 12 shows the same type figures as for model 1 with the same parameter configurations as in the main section, but with the exception of using a linear power-utilization curve (Figure 2 in place of the cubic (Figure 8). The optimization still aims to fill up workload to the threshold at 60% capacity, so as to make results comparable with the main model (which had its operational sweet spot at the same level).

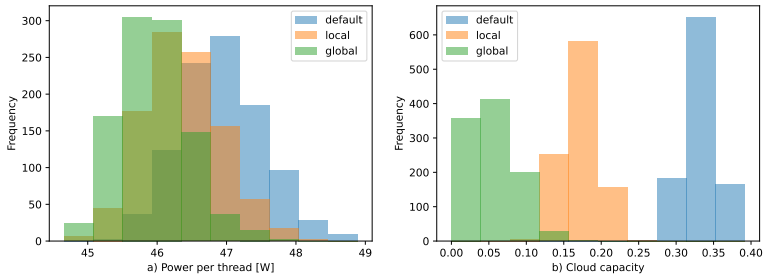

Figure 12: Results with alternative data center power utilization curve (linear in place of cubic), assuming a fixed edge capacity threshold at 60%

Note that for the linear power-utilization there is no longer a sweet spot below 100%. As the operational advantage of the sweet-spot is removed, we find that the per-thread power savings of the *local* and *global* modes are reduced, to 1.2% and 2.1% , respectively. For the linear curve, the sweet-spot is instead at full capacity. Table 6 shows the per thread savings for different thresholds.

Table 6: Operational savings for different packing threshold (linear power-utilization). In the absence of a sweet spot savings are much smaller.

	local	global
linear m=0.6	1.2%	2.1%
linear m=0.8	1.8%	2.5%
linear m=1.0	2.2%	2.5%

Table 7: Capacity savings for different packing thresholds (linear power-utilization). These numbers do not depend on the power utilisation curve.

	local	global
linear m=0.6	48.3%	83.8%
linear m=0.8	70.5%	99.9%
linear m=1.0	86.1%	99.9%

The claim on cloud capacity is unaffected by the choice of power utilization curve, which is why the right panel of Figure 12 is no different from Figure 10. Table 7 shows the reduction in capacity claim on the cloud for different packing targets.

Note that with a threshold at or above $m = 0.8$, that is, we pack until we are closer to full capacity at the edge the average claim on the cloud in this model is reduced by close to 100% (i.e. to zero). This depends on the parametrization of the joint load distribution, which in for this simulation had a left skew. For flatter workload generating distributions functions we see that this effect is much more moderated.

Effect of workload correlation. Figure 13 show the results from a simulation when the correlation parameters (of the Gaussian copula) was reduced from $\rho = 0.7$ to $\rho = 0.0$, that is if the dependence between the variables was relaxed such that a joint draw for the two workload types just corresponded to independent draws from their respective marginal distribution.

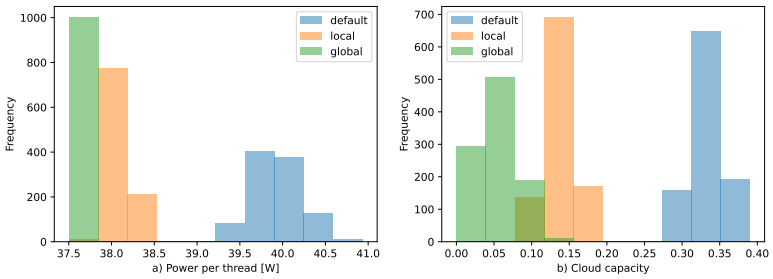

Figure 13: Effect of workload type correlation

We do not note any significant difference to the base model in the figures, which is also evident from the averages displayed in Tables 8 and 9 for power per thread and capacity savings, respectively. Thus we conclude that the model is not very sensitive to the correlation assumption.

Effect of skewed load distribution. Figure 14 show the results from a simulation in model 2 where the workload generating distribution function is replaced, such that flat (uniform) distribution is used in place of the original assumption of a left skew distribution.

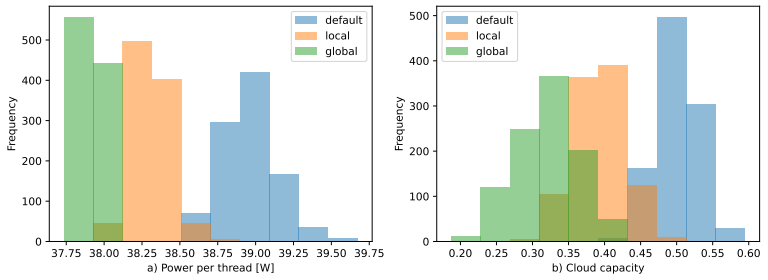

Figure 14: Effect of load distribution

We note a reduced distance between histograms for the different optimization strategies in both the left and right panel of the figure. The reduced savings under flat distribution assumption are also evident for averages in row 3 of Tables 8 and 9 for power-per-thread and capacity savings, respectively. Thus we conclude that the model have some sensitive to the distribution assumption.

Effect of time-dependant load distribution. Figure 15 show the results from a simulation where the time-homogeneous (and left skew) distribution assumed for the main model is replaced by a time-dependent distribution which alternates between a left skew and a right skew distribution.

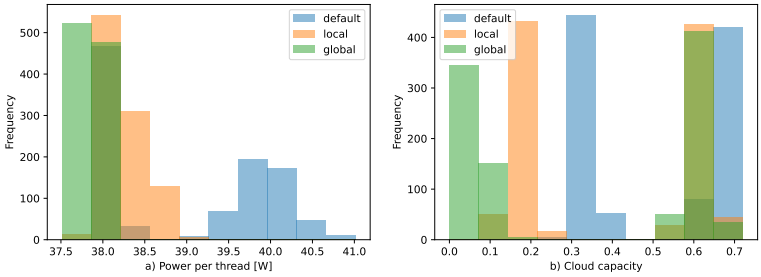

Figure 15: Effect of time-dependent load distribution

The histograms in the right side figure are multi-modal and difficult to interpret. Row 4 of Tables 8 and 9 for power-per-thread and capacity savings, respectively, shows reduced average savings. In fact the average savings are very similar in magnitude to what we saw for the flat distribution (which should perhaps be expected as the average of the left and right skew distribution is indeed flat). We conclude that the model is sensitive to the homogeneity assumption.

Table 8: Sensitivity of per thread savings.

	local	global
base model	3.7%	6.0%
- correlation	4.6%	5.9%
- distribution	1.6%	2.7%
- time dependence	2.1%	3.3%

Table 9: Sensitivity of capacity claim.

	local	global
base model	48.2%	83.6%
- correlation	58.6%	83.5%
- distribution	21.4%	35.9%
- time dependence	21.3%	33.5%