Exploring investment strategies for federated learning infrastructure in medical care

Ju Xing*, Xu Zhang†, Zexun Jiang‡, Ruilin Zhang‡, Cong Zha‡ and Hao Yin‡

*Department of Automation, Tsinghua University, †Department of Computer Science, University of Exeter,
‡Department of Computer Science, Tsinghua University

***Abstract*—Recently, federated learning has gained substantial attention in medical care where privacy-preserving cooperation among hospitals is required. However, in a real-world situation, the deployment of a federated learning system among hospitals requires heavy investment in computing and network infrastructure. Under such a case, making investment effective across computing power and network capability is essential. In this paper, we propose an investment methodology following the growth saturation of learning efficiency. We also systematically study the impacts of non-investment factors on the application of this methodology. With consideration of relevant cost models, the methodology is validated cost-effective.**

***Index Terms*—federated learning, investment, computing, network**

I. Introduction

Recent advances in the machine learning area flourishes diverse medical applications [1–3]. The sustained success of AI technologies heavily relies on high-quality data, which makes data sharing in medical care an even significant demand.

However, data security is a big concern blocking such sharing. Traditional medical care relies on the Regional Health Information Organization (RHIO) [4] for Electric Health Record (EHR) exchange between stakeholders of medical data. Although such centralized data aggregation paradigm is effective, it introduces an extremely high risk of data breaches. Besides, RHIO is hard to be implemented with large medical hospitals since they have more considerations on political feasibility and conflicts of interest. Recently, the emerging federated learning [5] [6] provides a new way for medical data cooperation among hospitals. With only the parameters exchanged, each hospital's data can be locked down in place, while the target model is trained with an entire feature space composed of all data. In medical care, federated learning typically falls into the cross-silo category since all the engaged participants are usually fixed across the training procedure.

Although federated learning has been widely applied to unleash the internal value of medical data [7–9] and is popular with industrial practice [10–12], the research against investment in its infrastructure lacks. Deploying a production-grade federated learning system in medical care requires heavy investment in the computing power inside hospitals and capability of the network outside hospitals: first, the current computing infrastructure of a hospital is typically CPU-intensive and storage-oriented. This kind of infrastructure is suitable for the business workload of on-premise information systems, such as Health Information System (HIS), Laboratory Information System (LIS), and Clinical Information Systems (CIS). However, to support massive federated learning tasks, specialized computing accelerators like GPUs must be used. From the standpoint of hospitals, they have a low desire to upgrade their existing infrastructures due to federated learning is neither a mature business nor a direction hospitals themselves willing to promote. Second, a hospital is usually a data silo and has little network traffic exchange with other entities like operators, cloud providers, and other hospitals. Accordingly, the network infrastructure connecting a hospital and the outside world is usually weak or even vacant. As a consequence, it is vital to manage the cost-effective investment on these two dimensions. In this work, computing power mainly refers to computing accelerators like GPU, while network capability mainly refers to network bandwidth.

Although computing power and network bandwidth determine the theoretical upper bound of learning efficiency, the growth of efficiency varies among couples of these two dimensions. Growth saturation usually leads to the critical points in the investment. In this work, we systematically analyze the learning efficiency and cost-effectiveness associated with investment through measurement. We disclose how the growth saturation of learning efficiency is influenced by continuous investment in computing power or network capability and explore how non-investment factors bias such influence. Besides, we evaluate the efficiency-cost trade-off of different investment strategies considering detailed cost models of computing power and network capability. We digest some crucial findings from our measurement here:

- The characteristics of workloads greatly impact the suitable proportions of computing power and network capability in the investment. A priori knowledge of workloads will contribute to a better investment decision.
- Scheduling algorithms with improvement for ”head-of-line” (HOL) blocking have a lower margin for investment benefit in computing power. Therefore, the investment in computing power should be conservative for hospitals where such kinds of schedules are taken.

J. Xing and X. Zhang are co-first authors.

- Investment strategies reaching efficiency growth saturation also achieves high cost-effectiveness. It indicates that investing along growth saturation curves gives a direction achieving the best trade-off.
- From a market perspective, the investment of network capability prefers the ordinary public network offered by operators, making the investment highly replicable across diverse operators.

II. Background

Federated learning is first raised by Google [13] [14] to deal with the cooperative model training among mobile clients. Due to the excellent trade-off between data privacy and model usability, federated learning becomes a popular research area. Federated learning includes various learning paradigm(e.g., horizontal, vertical, and transfer [5]) facilitating diverse learning settings. In a broader definition, some "multi-model" approaches like multi-task learning, meta-learning are also considered as federated learning. In this paper, we mainly focus on horizontal federated learning where the data from each hospital has a similar feature space.

Currently, the research in federated learning mainly focuses on algorithm optimization [15–17], communication efficiency [18] [19], and data privacy [20–22]. However, deploying a federated learning system in a real-world cross-silo setting requires investment in computing power inside silos and network interconnecting functional entities. Especially in medical care, the investment is even more important since this area's information infrastructure is a little bit lagging.

Although there have been some system works [23–25] and projects [26–29] targeted at federated learning; we have not seen any production-grade deployment of federated learning systems. Moreover, there exists little research work investigating the corresponding investment to our best knowledge.

III. Measurement methodology

A. Scenario

Figure 1 shows the typical architecture of a federated learning system in medical care. It is composed of a public cloud, edge clouds, and hospitals. A federated learning service is provided in the public cloud to facilitate the machine learning requirements of medical applications. Each hospital receives learning tasks from the service and uses local computing power for model training. Parameter servers are hosted in edge clouds for the aggregation of model parameters. The positions of edge clouds may vary from operators' near-hospital IDCs to base stations of cellular networks. Thus, there exist several choices for interconnections between the edge clusters and hospitals:

- Private network, where a tenant usually has exclusive network resources. A private network is usually constructed upon direct optical fiber connection or with advanced virtualization technologies (e.g., SD-WAN).
- Public network, where tenants share network resources. A public network is a common network category for Internet access of families and companies. Both private and public networks here refer to the class of fixed network access.
- Cellular network. With emerging wireless communication technologies like 5G or 6G, the interconnection between a hospital and edge cloud may directly use a cellular network. In this work, we use the 5G network as a representative.

B. Problems to be explored

1. How does the growth saturation of learning efficiency change with continuous investment in computing power and network?

Given working sets, continuous investment in these two dimensions will contribute to learning efficiency. However, the growth speed may gradually converge due to a mismatch between computing power and network capability. In other words, over-subscription or waste in either dimension will make growth and investment inefficient. Thus exploring how computing power and network capability interact with each other influencing the growth saturation is essential.

2. To what extent this change is biased by non-investment factors, such as workload characteristics and scheduling algorithms of learning tasks?

The investment only decides the capacity for data computation and transmission. The actual utilization of computing power and network capability mainly depends on the combination of workload itself and software implementations. In our case, we choose workload characteristics and scheduling algorithms as typical representatives.

3. Is the investment around growth saturation economic? Furthermore, how do cost models of investment dimensions influence the investment strategies?

From the aspect of cost-effectiveness, the investment achieving the highest learning efficiency is usually not good enough. The investments following growth saturation will also have divergence from the optimal one. We want to measure this divergence to validate its benefits. Besides, the cost models of investment dimensions directly affect the resulting cost-effectiveness. Compared with computing power, the cost models of network capability typically vary due to operators' charging modes against network categories. Explore the impacts brought by this variety will lead us to a more accurate investment target.

C. Method

To avoid ambiguity, we explain two concepts here:

- **Learning efficiency**: In our scenario, we assume that tasks are arrived in batches and use the average completion time T of tasks as a metric to measure the efficiency.
- **Learning cost-effectiveness**: We use the efficiency brought with per unit of investment amount($\frac{1}{T \cdot c}$, where c is the amount of investment) as a metric to measure the cost-effectiveness.

In order to achieve high efficiency, the investment should balance between the computing power and the network capability, thus neither dimension is over-subscribed or wasted.

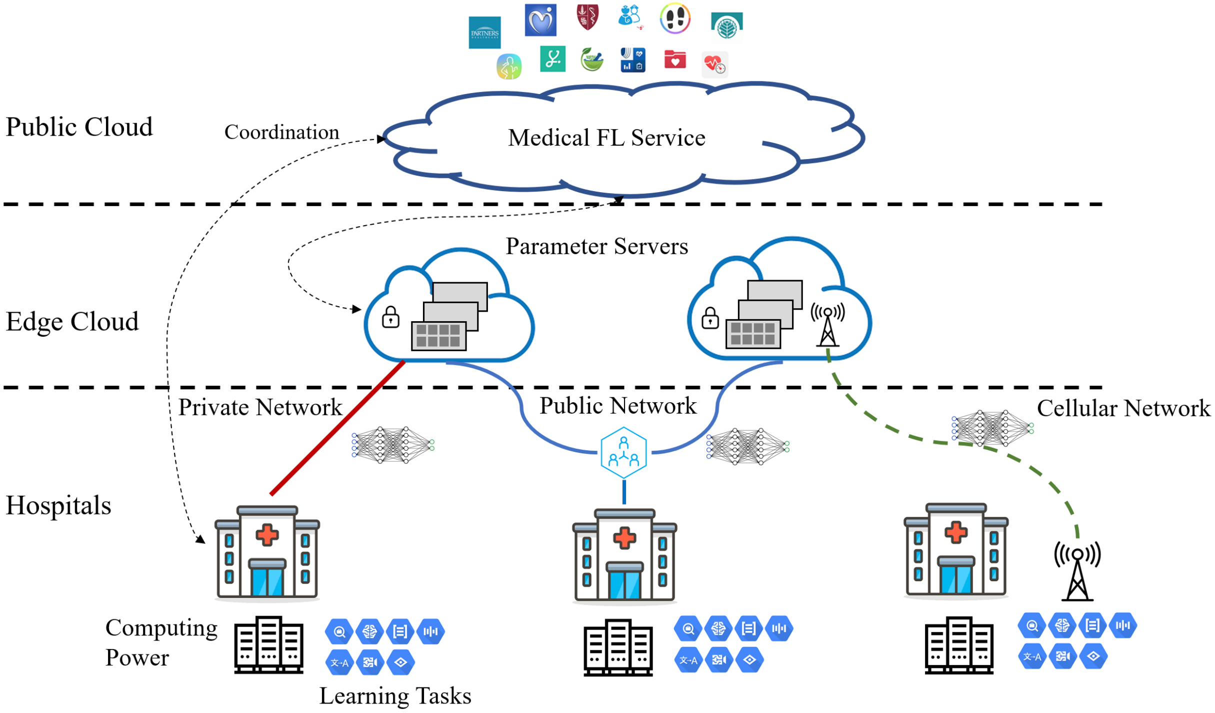


Figure 1: The architecture of a federated learning system in medical care.

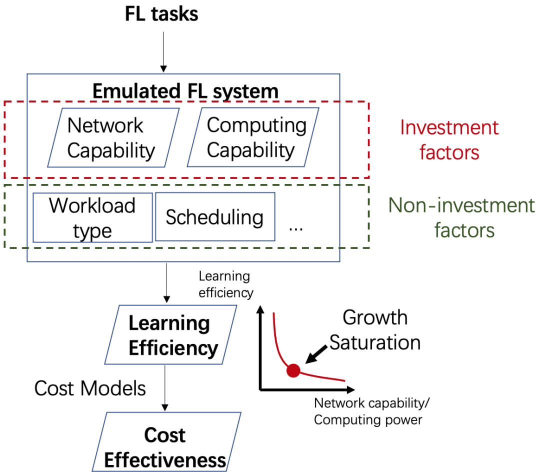


Figure 2: Exploration Method.

Figure 2 shows the method we follows: By using an emulated federated learning system, we first measure the learning efficiency with target tasks under different investments on computing power and network capability. With continuous investment in either of these two dimensions, the growth of learning efficiency tends to saturate. We assume such saturation represents the fore-mentioned ”balanced state” from the perspective of efficiency and give some formal definitions related with the state:

Definition 1 (Saturated Investment Point)**.** A Saturated Investment Point (SIP) designates an investment dimension value (There are computer power's SIPs and network capability's SIPs). Once the value of the investment dimension exceeds this point, the growth rate of the learning efficiency declines below a specified threshold t.

Definition 2 (Saturated Investment Curve)**.** The Saturated Investment Curve (SIC) of an investment dimension is composed of SIPs with corresponding to the consecutive sampling in the other dimension. For example, the SIC of computing power is composed of SIPs whose network capability is consecutively sampled.

A SIP typically represents a state where the effectiveness brought with investment starts saturating. In this work, we empirically set threshold t as 5%. The SIC indicates how computing power and network capability interact with each other affecting the growth saturation. Besides, through the observation of SICs under non-investment factors, we can infer how those factors bias this affection on growth saturation. Furthermore, we adjust non-investment factors to explore their affections on the SICs of investment dimensions. We also define a metric to roughly estimate ”balanced states” from the workload's perspective. Finally, we evaluate the cost-effectiveness of investments with consideration of cost.

Definition 3 (Balance Ratio)**.** Balance Ratio (BR) is a demand ratio of computing power and network capability derived from workload only. It is calculated by the equation $BR = \frac{\theta}{M \cdot \lambda}$. θ is the total GPU time a batch of tasks consuming in an epoch, and M is the total model size of these tasks. λ is the compression ratio for model parameters.

IV. System setting

Since there is no production-grade federated learning system in medical care for measurement, we implement an emulation system using NS-3 [30], a discrete-event network emulator for distributed systems.

A. Workload

We synthesize the workloads from some open-sourced medical applications. Table I lists applications with detailed informations. These applications are classified into three categories based on the model size in the learning task: group I and group II contain applications with medium or large model sizes, while group III contains that with small model sizes but a bit more epochs. We mix these applications with various ratios to synthesize workloads (see Table II). Since our research mainly focuses on the investment in network and computing power, we assume that epochs we introduced for workloads have made learning accuracy reach a certain level. Besides, we associate a **parallelism degree** for each learning task to reflect the distributed paradigm of training inside a hospital. The parallelism degree is set to 1 or 2 randomly.

B. Network capability

Table III illustrates typical statistics we used for networks we introduced. The private network usually has high-performance interconnecting and few variances in network conditions. Besides, identical upstream/downstream bandwidth is a significant feature. The public network is a common network condition provided by operators for fixed access of Internet and usually has un-equal upstream/downstream bandwidth. 5G network has medium latency while the downstream bandwidth is exceptionally high. The fixed downstream bandwidth is consistent with the 5G service provided by operators currently. Since both public networks and 5G networks share bandwidth with multi-tenants, we use **Bandwidth Utility** to describe the average available bandwidth. In our case, this value is set to be 90%.

C. Computing power

We fix GPU type to NVIDIA Quadro P4000 with 8 GiB memory and count GPU quantity to measure the investment. The GPU time of a workload is collected with execution using one such GPU card. The later-introduced cost model of computing power is also calculated with this kind of GPU. Although this strawman solution may not represent any classical settings, it sufficiently help with the exploration of our problems, and the rules to be disclosed will not be biased.

D. Cost model

We model the cost of network capability and computing power with a per-month granularity. We introduce **Service life** l (unit: month) to denote the period during which GPU functions without any performance degradation. Thus the cost model of computing power is $C = n * \frac{p}{l}$, where p denotes single GPU cost and n denotes the GPU quantity. In our emulation, we set l as 14 and uses the price of Quadro P4000 from Amazon as a reference. As a result, the cost model for computing power is $p = 50 * n$ dollars per month.

However, the cost models of network capability diverse either among operators, and related to network categories. We choose China Telecom as a representative and uses its charging modes in Shanghai [31] [32] as a baseline to derive the cost models in tableIV. For the private network, the cost has a linear relationship with bandwidth due to the exclusivity of network resources in this case; For the public network, this relationship turns into a piece-wise linear fashion. With bandwidth increase, the cost of per unit bandwidth decreases quickly; For the 5G network, the cost is mainly measured in traffic and is not relevant to network bandwidth. The reason causes this phenomenon is that the commercialization of 5G is at a very early stage, and operators rely on this charging mode to compensate for investment in 5G infrastructures(e.g., base stations).

E. Other factors

Task scheduling. Here, we mainly focus on the scheduling algorithms of learning tasks inside each hospital. The aggregation scheduling in the edge cloud is out of scope. Three algorithms are considered:

- **Exclusive**: Each task locks its GPUs exclusively until finishing all the computation;
- **Time-sharing**: When a task sends its parameters for aggregation, the GPUs it takes will be released for other tasks' use;
- **Time-sharing w/ priority**: This scheduling algorithm is similar to Time-sharing, except that short tasks have higher priority to be scheduled in available GPUs. All tasks are sorted by the GPU time in an epoch.

The speed of context switch in GPU is set as 30 gigabytes per second.

Communication efficiency. We use randomized quantization [33] for parameter compressions. The quantization level is 8, and the quantization bucket we choose is 512. We set the compression/decompression speed as 2 milliseconds for each megabyte of data in the GPU [34].

Parameter aggregation. We simulate the SGX-based parameter server [35]. A pair-wise aggregation of one megabyte of data takes about 1.5 milliseconds, and the aggregation happens every 2 epochs. Besides, we consider the encryption and decryption for exchanged parameters and set corresponding speeds as 11.2 milliseconds and 13.2 milliseconds for each megabyte of data in the CPU [36].

V. Evaluation

A. Efficiency growth with investment

Figure 3 illustrates the efficiency surface with discrete sampling on network bandwidth and GPU quantity. We got all ACTs via the emulation under the private network setting. The number of engaged hospitals is 20, and the interconnections between aggregation edge cloud and hospitals are private networks. The workload and scheduling algorithm we used are workload II and Time-sharing, respectively. The

Table I: Medical applications.

Group	Index	Application	Model	Model size(MB)	Epoch	GPU time(s/epoch)
Group I	A	Biomedical image segementation	Unet	355.13	10	39.33
	B	retinal Blood vessel segmentation	BCDU-RBVS	236.00	50	0.57
	C	Skin lesion segmentation	BCDU-LS	242.10	10	77.83
	D	Contextual representation for clinical notes	ClinicalBERT	417.69	10	1837.08
Group II	E	detection of COVID-19 cases	COVIDNECT-2-L	4280.32	20	507.90
	F	detection of COVID-19 cases	COVIDNECT-2-S	1986.56	20	392.57
	G	detection of COVID-19 cases	COVIDNECT-2-L RAD	4280.32	20	518.83
Group III	H	Medical entity recognition	BiLSTM-CRF	20.80	100	146.21
	I	Medical diagnosis	MD-RNN	6.55	100	512.90
	J	Medical diagnosis	MD-RCNN	18.10	200	540.16
	K	Representations of medical concepts	GRAM	7.19	100	4.99

[1] All GPU times are measured with a GPU card of Nvidia Quadro.

Table II: Workloads.

Workload	Group I	Group II	Group III
Workload I	10%	80%	10%
Workload II	10%	10%	80%
Workload III	30%	40%	30%

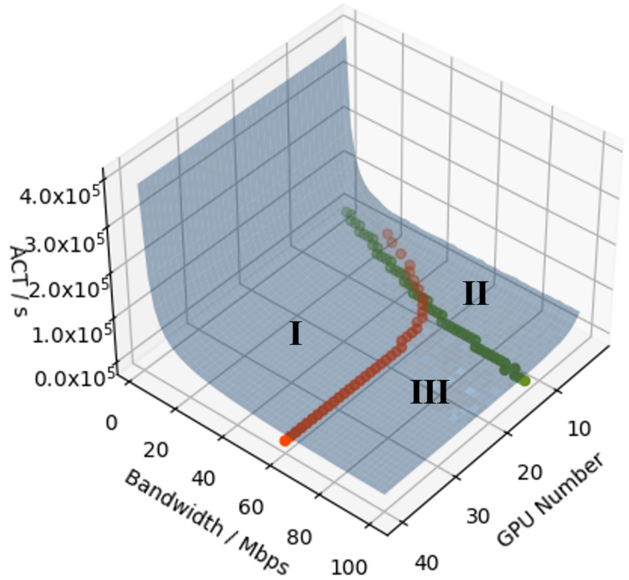


Figure 3: Efficiency surface and SICs.

surface descends with more GPUs or network bandwidth, and meanwhile, gradually becomes flat. Therefore, the growth of efficiency indeed converges when efficiency itself gets high.

The SICs of computing power and network capability are drawn in the figure to denotes the converging positions. These two SICs splits the investment plane, which is constructed by computing power and network capability, into three primary regions. When the investment falls into region I, it represents an excessive investment on computing power while an insufficient investment on the network capability; When the investment falls into region II, the situation is on the contrary. III represents an excessive investment in both network and computing power. The learning efficiency exhibits similar characteristics with other network settings and non-investment factors. Therefore, there is a optimal choice(point of intersection of two SICs) for the investment in computing power and network capability from the aspect of margin benefit.

B. Impacts of non-investment factors on SICs

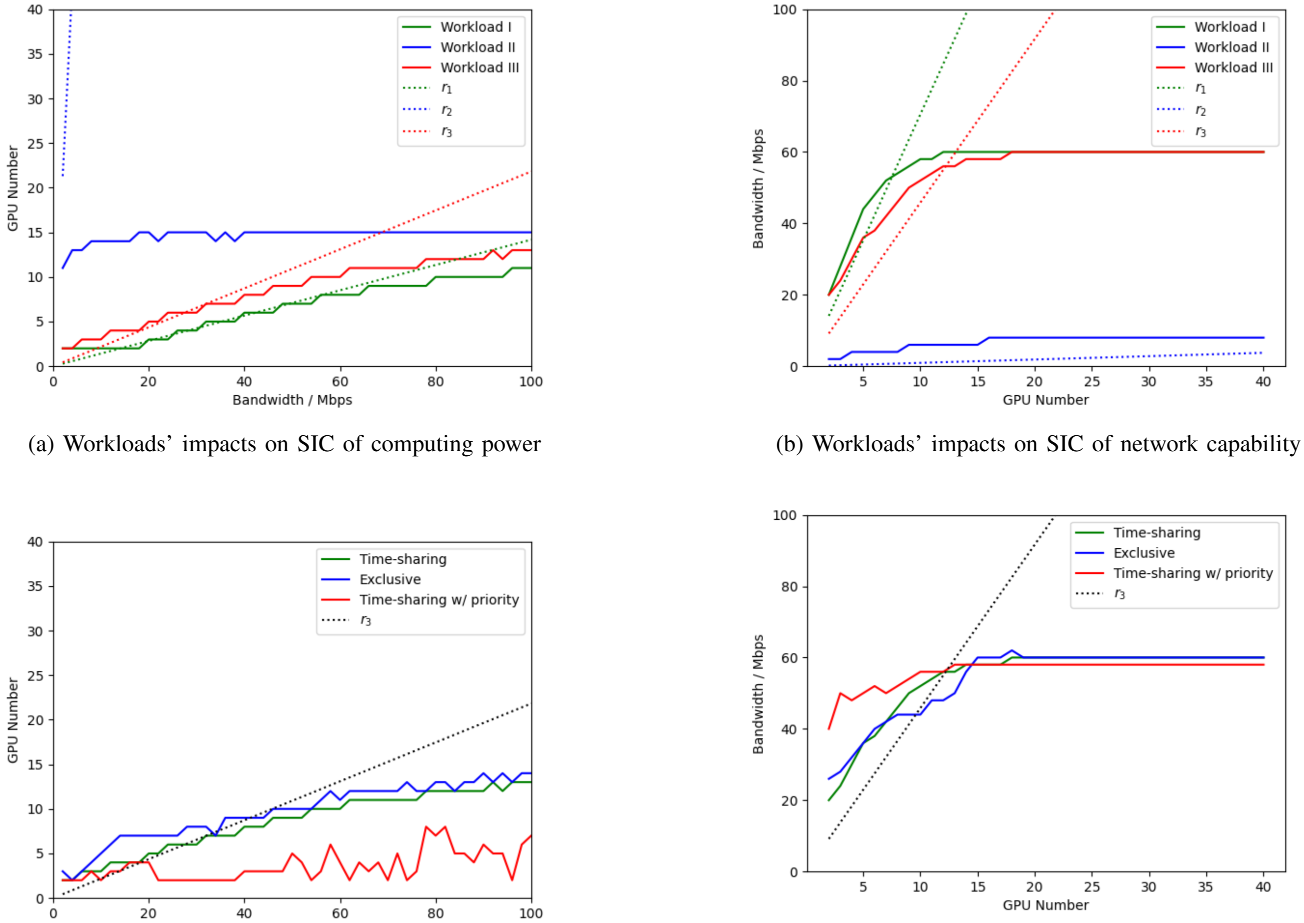
Figure 4 shows how non-investment factors influence the SICs of computing power and network capability. The number of engaged hospitals is 20, and the interconnections between aggregation edge cloud and hospitals are private networks. The scheduling algorithm used in Figure 4b and Figure 4a is ”Time-sharing”, the workload used in Figure 4a and Figure (4d) is workload III.

Figure 4a shows that SICs of computing power tend to converge when network capability continuously increases. Once the network capability reaches some degree, it is a surplus resource compared with computing power; thus the SIP of computing power will not be steered by network capability. The SIC with workload II converges most quickly since the communication overhead of this workload is relatively low ,and network capability gets ”superfluous” easily. Curves r_1, r_2, and r_3 reflects balance ratios of workloads. These ratios are rough estimations of demand balance between computing power and network capability, while SICs reflect this balance more accurately. The figure shows these two kinds of curves are consistent in average slope before SICs converge.

Figure 4b shows the SICs of network capability under different workload. Similar to the case of computing power, the surplus of network capability drives SICs to converge. However, high heterogeneity of workloads slows down the convergence. Workload III demands the least computing power while the SIC with it converges most slowly. Workload I and workload II both have a principle category in the workload. Therefore, the SIC with workload I converges fast due to its lower demand for computing power in an epoch.

Table III: Network conditions.

Network Category	Latency(ms)	Bandwidth(Mbps)	Up/down Ratio	Bandwidth Utility(%)	Jitter(ms)	Packet Loss(%)
Private Network	10	customized	1	100%	0.05	0
Public Network	25	customized	0.1	90%	10	0
5G Network	15	1000	0.1	90%	5	0



(a) Workloads' impacts on SIC of computing power

(b) Workloads' impacts on SIC of network capability

(c) Scheduling algorithms' impacts on SIC of computing power

(d) Scheduling algorithms' impacts on SIC of network capability

Figure 4: Non-investment factors' impacts on SICs.

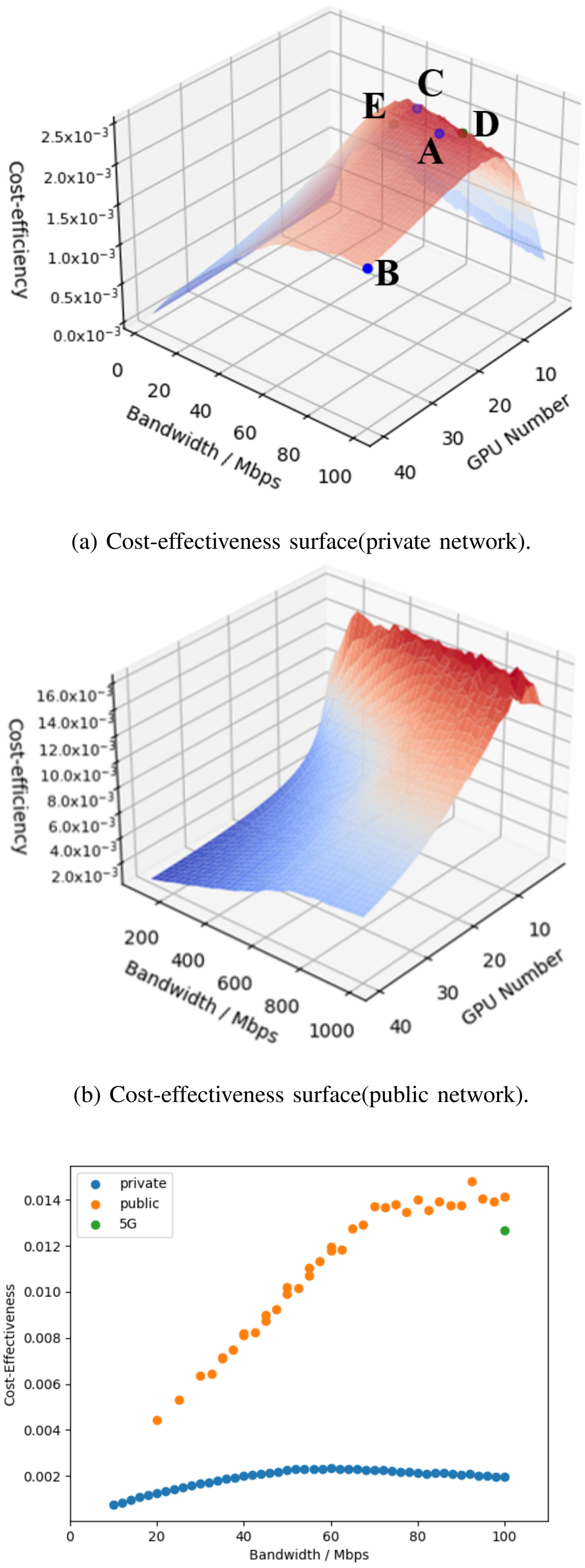
Table IV: Cost models of network capability.

Category	Cost($/month)
Private Network	$C = 1200 + (B - 10) * 23$
Public Network	$C = \begin{cases} 15.47 + \frac{6.41(B-100)}{100} & B < 300 \\ 28.46 + \frac{6.41(B-300)}{200} & 300 < B < 500 \\ 33.57 + \frac{6.41(B-500)}{500} & 500 < B \end{cases}$
5G Network	$0.152 * T$

[1] B denotes the network bandwidth(Mbps), T denotes the network traffics(GB)

Figure 4c shows that the SICs of computing power under **Time-sharing** and **Exclusive** scheduling are close and tend to converge, although the previous one often results in higher GPU utilization. Nevertheless, the SIC under **Time-sharing w/ priority** oscillates at lower values. On the one hand, this scheduling algorithm skews the task distribution, which is highly beneficial for our definition of efficiency. Therefore, the absolute value of SIP is low; on the other hand, oscillation is caused by the execution of long tasks when the GPU number is small.

The SICs of network capability under **Time-sharing** and **Exclusive** scheduling are still close but smoother in the Figure 4d. The SIC under **Time-sharing w/ priority** is lower than the other two curves. With this algorithm, an increase in network capability mainly helps improve the efficiency of short tasks and soon can not contribute to the improvement of overall efficiency anymore. Furthermore, this algorithm makes



(a) Cost-effectiveness surface(private network).

(b) Cost-effectiveness surface(public network).

(c) Cost-effectiveness vs. network categories.

Figure 5: Cost-effectiveness of investment.

computing power easier to be surplus, which is reflected by a faster convergence of corresponding SIC. The balance ratio curves are also illustrated in Figure 4c and Figure 4d for comparison. The scheduling algorithms can hardly shift the balance demand in reality far away from that disclosed by workload. It is noted that in the public network setting, the results are similar.

C. Cost-effectiveness Analysis

The validity of investment along SICs. Figure 5a illustrates the surface reflecting cost-effectiveness of investment with the same setting as that in the efficiency case. Since the per-unit cost of computing power and network capability diverges(3.8$ vs. 88$), the resulting surface forms a peak, where investment results in the best cost-effectiveness(Point A). Point B represents the investment achieving the highest efficiency. It has a loss of 21.89% in cost-effectiveness compared with optimal value. Point C is the intersection point of two SICs in the Figure 3 with a loss of 3.17%in cost-effectiveness and 31.78% in efficiency; Point D is the projection of B on the SIC of computing power. It has a loss of 6.99% in cost-effectiveness and 31.78% in efficiency; Point E is the projection of B on the SIC of network capability. It has a loss of 4.18% in cost-effectiveness and 23.99% in efficiency. This indicates that investing along SICs achieves high margin benefit, and results in high cost-effectiveness with trading some performance(usually no more than 30%). In the public network setting, this pattern still holds while the peak of the surface is closer to small GPU values due to a cheaper cost for network bandwidth (Figure 5b). The bandwidth in the figure refers to the downstream one.

Impacts of network category on cost-effectiveness. Figure 5c shows the changing of cost-effectiveness with the network capability. The computing power is fixed as 10 GPUs. The cost-effectiveness in the private network slightly rises up and descends after. While in the public network, the cost-effectiveness continuously rises with the increase in network capability. This continued growth in cost-effectiveness is caused by descending price per unit of bandwidth under public network. In other words, the public network should be chosen as a specific target for the investment in network capability. Besides, the cost-effectiveness in the 5G network is only a point due to fixed bandwidth in this network category.

VI. Conclusion

In this work, we comprehensively explore the efficiency and cost of federated learning investment in the medical care scenario. The investment following SICs achieves good trade-offs between efficiency and cost. SICs can be largely influenced by non-investment factors like workload characteristics and task scheduling algorithms, which investors should consider to make better decisions. At the same time, public network will be a better choice for investment in network capability due to its price advantages.

ACKNOWLEDGMENT

This research work has received key funding from The National Key Research and Development Program of China (2018YFB2100804), National Natural Science Foundation of China (Grant No. 92067206, No. 61972222 and No. 61902178), Hunan Provincial Research and Development Program in Key Areas (2019WK2071), Natural Science Foundation of Jiangsu (BK20190295), Leading Technology of Jiangsu Basic Research Plan (BK20192003), and European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 898588.

REFERENCES

[1] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164, 2018.

[4] V. N. Patel, R. V. Dhopeshwarkar, A. Edwards, Y. Barrón, J. Sparenborg, and R. Kaushal, "Consumer support for health information exchange and personal health records: a regional health information organization survey," *Journal of medical systems*, vol. 36, no. 3, pp. 1043–1052, 2012.

[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[7] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[8] S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 270–274.

[9] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, and X. Jiang, "Privacy-preserving patient similarity learning in a federated environment: development and analysis," *JMIR medical informatics*, vol. 6, no. 2, p. e20, 2018.

[10] IBM, "Ibm federated learning," https://ibmfl.mybluemix.net/.

[11] Owkin, "Owkin," https://owkin.com/.

[12] sherpa.ai, "sherpa.ai," https://sherpa.ai/.

[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[14] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[15] J. Wang, A. K. Sahu, Z. Yang, G. Joshi, and S. Kar, "Matcha: Speeding up decentralized sgd via matching decomposition sampling," in *2019 Sixth Indian Control Conference (ICC)*. IEEE, 2019, pp. 299–300.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[17] X. Li, W. Yang, S. Wang, and Z. Zhang, "Communication efficient decentralized training with multiple local updates," *arXiv preprint arXiv:1910.09126*, vol. 5, 2019.

[18] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

[19] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," *arXiv preprint arXiv:2006.10672*, 2020.

[20] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[21] A. Segal, A. Marcedone, B. Kreuter, D. Ramage, H. B. McMahan, K. Seth, K. Bonawitz, S. Patel, and V. Ivanov, "Practical secure aggregation for privacy-preserving machine learning," 2017.

[22] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning," in *2020 {USENIX} Annual Technical Conference ({USENIX}{ATC} 20)*, 2020, pp. 493–506.

[23] Z. Chai, A. Ali, S. Zawad, S. Truex, A. Anwar, N. Baracaldo, Y. Zhou, H. Ludwig, F. Yan, and Y. Cheng, "Tifl: A tier-based federated learning system," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, 2020, pp. 125–136.

[24] G. Damaskinos, R. Guerraoui, A.-M. Kermarrec, V. Nitu, R. Patra, and F. Taiani, "Fleet: Online federated learning via staleness awareness and performance prediction," in *Proceedings of the 21st International Middleware Conference*, 2020, pp. 163–177.

[25] C. He, S. Li, J. So, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, L. Shen *et al.*, "Fedml: A research library and benchmark for federated machine learning," *arXiv preprint arXiv:2007.13518*, 2020.

[26] T. F. Authors, "Federated ai technology enabler," https://www.fedai.org/, 2021.

[27] T. P. Authors, "Paddlefl," https://github.com/PaddlePaddle/PaddleFL/, 2021.

[28] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *arXiv preprint arXiv:1811.04017*, 2018.

[29] T. T. Authors, "Tensorflow federated," https://www.tensorflow.org/federated/, 2021.

[30] T. ns 3 developers, "Ns-3 simulator," https://github.com/nsnam/.

[31] "Public network pricing," https://sh.189.cn/.

[32] "5g network pricing," https://sh.189.cn/.

[33] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1709–1720, 2017.

[34] xinyandai, "Quantization schema," https://github.com/xinyandai/gradient-quantization/.

[35] J. Xing, Z. X. Jiang, and H. Yin, "Jupiter: A modern federated learning platform for regional medical care," in *2020 IEEE International Conference on Joint Cloud Computing*. IEEE, 2020, pp. 21–21.

[36] T. I. developers, "Aes-ni performance testing," https://software.intel.com/content/www/us/en/develop/articles/intel-aes-ni-performance-testing-on-linuxjava-stack.html/.