Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences: 58(S): 59-64 (2021)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(58-sp1)733

Pakistan Academy of Sciences

Research Article

Custom Built of Smart Computing Platform for Supporting Optimization Methods and Artificial Intelligence Research

Indar Sugiarto[1*], Doddy Prayogo[2], Henry Palit[3], Felix Pasila[1], Resmana Lim[1], Agustinus Noertjahyana[3], I Gede Widyadana[4], Surya Hermawan[2], Agustinus Bimo Gumelar[5], and Bernardo Nugroho Yahya[6]

[1]Department of Electrical Engineering, Petra Christian University, Jl. Siwalankerto No.121-131, Surabaya, 60236, Indonesia
[2]Department of Civil Engineering, Petra Christian University Surabaya, 60236, Indonesia
[3]Department of Informatics, Petra Christian University, Surabaya, 60236, Indonesia
[4]Department of Industrial Engineering, Petra Christian University Surabaya, 60236, Indonesia
[5]Department of Information System, Narotama University, Jl. Arief Rachman Hakim 51, Sukolilo, Surabaya, 60117, Indonesia
[6]Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, 107, Imun-ro, Dongdaemun-gu, Seoul, 130-791, Korea

Abstract: This paper describes a prototype of a computing platform dedicated to artificial intelligence explorations. The platform, dubbed as PakCarik, is essentially a high throughput computing platform with GPU (graphics processing units) acceleration. PakCarik is an Indonesian acronym for *Platform Komputasi Cerdas Ramah Industri Kreatif,* which can be translated as "Creative Industry friendly Intelligence Computing Platform". This platform aims to provide complete development and production environment for AI-based projects, especially to those that rely on machine learning and multiobjective optimization paradigms. The method for constructing PakCarik was based on a computer hardware assembling technique that uses commercial off-the-shelf hardware and was tested on several AI-related application scenarios. The testing methods in this experiment include: high-performance lapack (HPL) benchmarking, message passing interface (MPI) benchmarking, and TensorFlow (TF) benchmarking. From the experiment, the authors can observe that PakCarik's performance is quite similar to the commonly used cloud computing services such as Google Compute Engine and Amazon EC2, even though falls a bit behind the dedicated AI platform such as Nvidia DGX-1 used in the benchmarking experiment. Its maximum computing performance was measured at 326 Gflops. The authors conclude that PakCarik is ready to be deployed in real-world applications and it can be made even more powerful by adding more GPU cards in it.

Keywords: Artificial Intelligence, Machine Learning, Multi-objective Optimization, Graphics Processing Unit Accelerator, High Throughput Computing.

1. INTRODUCTION

In recent years, interest in Artificial Intelligence (AI) researches is increasing and showing its fruitful results in many areas including creative industry sectors. With the advent of Industrial Revolution 4.0, the need of implementing practical but robust AI becomes more and more demanding. Thus, researches in this area are blooming [1, 2]. However, running high-impact AI researches especially on Deep Learning (DL) requires a high-performance computing platform [3, 4]. This is an inevitable consequence since DL, like other machine learning approaches, usually works on massive data to automatically gain its working parameters.

Several dedicated computing platforms have

Received: April 2021; Accepted: July 2021
*Corresponding Author: Indar Sugiarto <indi@petra.ac.id>

been produced by diverse vendors that target this ever-growing research field [5]. For example, Nvidia Corporate has several products that are capable of high-performance computing up to several hundred teraflops, scaled from a workstation level to a server-class [6, 7]. However, those platforms are either very expensive or a part of a cloud computing service that does not give any direct means of maintenance for the AI researcher. For example, the price for an entry-level one Nvidia DGX Station is USD 149 000, which is way above standard research grants in Indonesia (around USD 3 500 yr^{-1}). In this circumstance, the authors have developed a smart computing platform that can be built to support AI researches. The platform, dubbed as PakCarik, is essentially a high throughput computing platform with GPU (graphics processing unit) acceleration. PakCarik is an Indonesian acronym for *Platform Komputasi Cerdas Ramah Industri Kreatif*, which can be translated as "Creative Industry friendly Intelligence Computing Platform". This paper describes how PakCarik was built and tested on several application scenarios.

2. MATERIALS AND METHODS

PakCarik aims to provide a complete development and production environment for AI-based projects, especially those that rely on optimization and machine learning paradigms. This research area is well-known for its challenging but intriguing methods for uncovering hidden information in massive unstructured data [8].

As a platform targeting industrial applications, PakCarik is equipped with various open-source libraries that enable the developer to quickly develop and deploy their projects. Special messaging service software is also installed in PakCarik, making it a complete framework for developing an IoT-based system. This software will act as an integrated IoT broker inside PakCarik that accommodates various messaging protocols such as MQTT and Kafka. The platform was designed such that the connection among different protocols is seamless. Such mechanisms are known to be challenging but very useful for developing complex IoT applications [9]. Figure 1 shows how PakCarik will be deployed as an IoT-based platform.

From a hardware perspective, PakCarik is a high throughput computer that is built using COTS (commercially off-the-shelf) components. Using COTS, the cluster can be customized to meet the customer budget whilst achieving high performance in a self-maintainable fashion [10]. Currently, PakCarik has two prototypes: PC1 and PC2. Table 1 shows the difference between PC1 and PC2.

3. RESULTS AND DISCUSSION

PakCarik has undergone these preliminary tests: high-performance lapack (HPL) benchmarking, message passing interface (MPI) benchmarking, and TensorFlow (TF) benchmarking. Table 4 shows the benchmarking results on PakCarik using the only CPU without GPU acceleration.

From the experiments, this research gained some insight into the performance of PakCarik at normal speed (without overclocking). Although the two prototypes of PakCarik use different processors, the performance is not much different. This opens the possibility of combining both prototypes into one cluster. Currently, PakCarik is undergoing thorough tests in Turbo mode where researchers overclock the processors up to 5 GHz (or more). However, we haven't found the highest stable operating frequency since these tests require the presence of an advanced cooling system (only one of the prototypes, which is PC1, that has a liquid cooling system at the moment).

This paper also performed a benchmarking for the GPU accelerator with Deep Learning image processing applications. For this, the authors use two models: inception-3 [11] and restnet-50 [12]. The comparison was made by measuring how many images were processed during the training phase since this phase usually takes more resources and a longer time compared to the inference phase. In these models, the batch size for the training was set to 32. The results of these experiments were compared to the results of different machines whose data are available online [13]. Figure 2 and Figure 3 show the performance of PakCarik compared to the performance of other platforms.

The purpose of the experiment shown in Figure 2 and Table 2 was to compare head-to-head of this platform (PakCarik) to the other platforms that were restricted to use only a single GPU in the

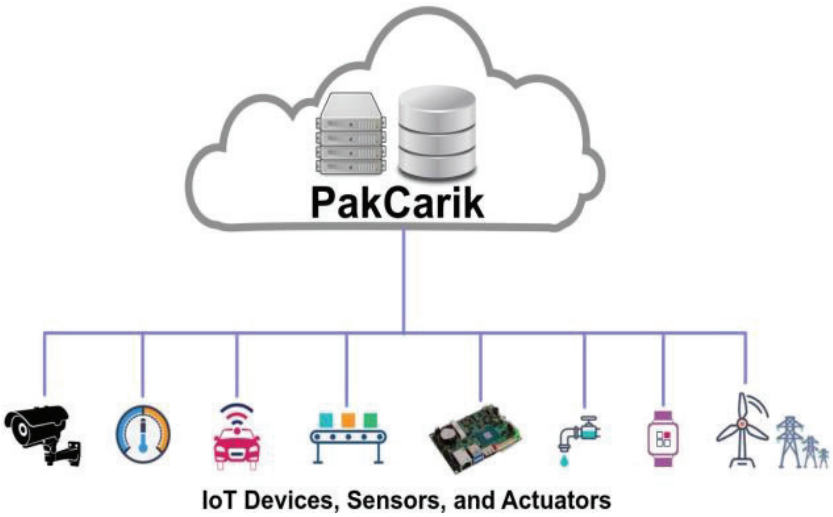


Fig. 1. IoT topology based on the PakCarik ecosystem. In this scenario, external IoT gateways are not necessary since they are embedded in PakCarik.

Table 1. Two protypes of PakCarik.

Component	PC1	PC2
Processor	Intel i9-7900X (10-cores) @3.30GHz	AMD Threadripper 2990WX (32-cores) @3.00GHz
Memory	40GB DDR4	32GB DDR4
Storage	500GB SSD SATA3	250GB SSD NVMe/PCIe
GPU	GTX 1070ti (2432 cores, 8GB DDR5)	RTX 2080ti (4352 cores, 11GB DDR5)
PSU	1 000 W	1 200 W
Form factor	ATX-Tower	Rackmount 4U

Table 2. Platforms with a single GPU accelerator for the experiment.

Platform	Description
Platform-1	NVIDIA® DGX-1 with 1-GPU Tesla® P100
Platform-2	Google Compute Engine with 1-GPU NVIDIA® Tesla® K80
Platform-3	Amazon EC2 with 1-GPU NVIDIA® Tesla® K80
Platform-4	Dell Gaming Laptop G7 with 1-GPU GTX-1050ti
Platform-5	PC1 (PakCarik prototype 1), see Table 1
Platform-6	PC2 (PakCarik prototype 2), see Table 1

Table 3. Platforms with multiple GPU accelerators were used for the experiment.

Platform	Description
Platform-1	NVIDIA® DGX-1 with 8-GPU Tesla® P100
Platform-2	Google Compute Engine with 8-GPU NVIDIA® Tesla® K80
Platform-3	Amazon EC2 with 8-GPU NVIDIA® Tesla® K80
Platform-4	Dell Gaming Laptop G7 with 1-GPU GTX-1050ti
Platform-5	PC1 (PakCarik prototype 1), see Table 1
Platform-6	PC2 (PakCarik prototype 2), see Table 1

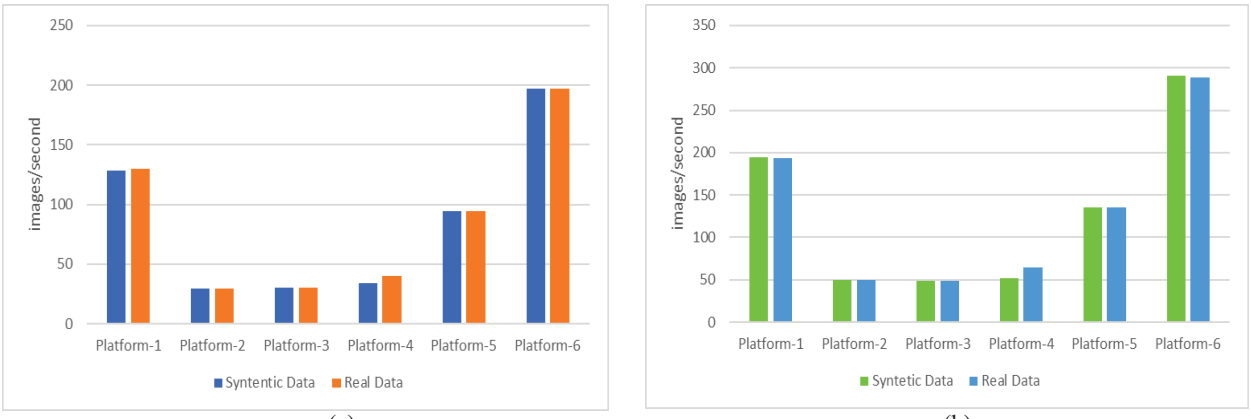
processing of Inception v3 and Restnet-50. This needs to be done since each prototype of PakCarik only has one GPU. The results show that the second prototype of PakCarik (PC2) outperforms the other platforms. This happens because PC2 uses a GPU card that has more cores and memory compared to the other platforms. Platform-2 and Platform-3, which are commonly used by cloud computing communities and when restricted only to using a single GPU, are quite similar to Platform-4, which is very common to be found in nowadays computer market.

The purpose of the experiment shown in Figure 3 and Table 3 was to compare head-to-head of this platform (PakCarik) to the other platforms with their maximum configuration. Platform-1 can have up to eight GPU cards, whereas Platform-2 and Platform-3 can have more cards but were restricted to use only eight cards in the experiment. On the other hand, PakCarik has a single GPU card in each prototype. The results show that Platform-1 (NVIDIA DGX-1) with eight GPU cards outperforms the other platforms.

This is a predicted result since more cards mean more computing power [14, 15]. However, even though PC2 has a single GPU card, its performance

Table 4. Result of the preliminary experiments on PakCarik.

Experiment	PC1	PC2
HPL	326 Gflops	317 Gflops
MPI	2.76 s	2.85 s
TF	133.96 fps	133.07 fps



(a) (b)

Fig. 2. Performance comparison when running (a) Inception v3 model and (b) Restnet-50 model using platforms described in Table 2. The higher the value (images s^{-1}) the better.

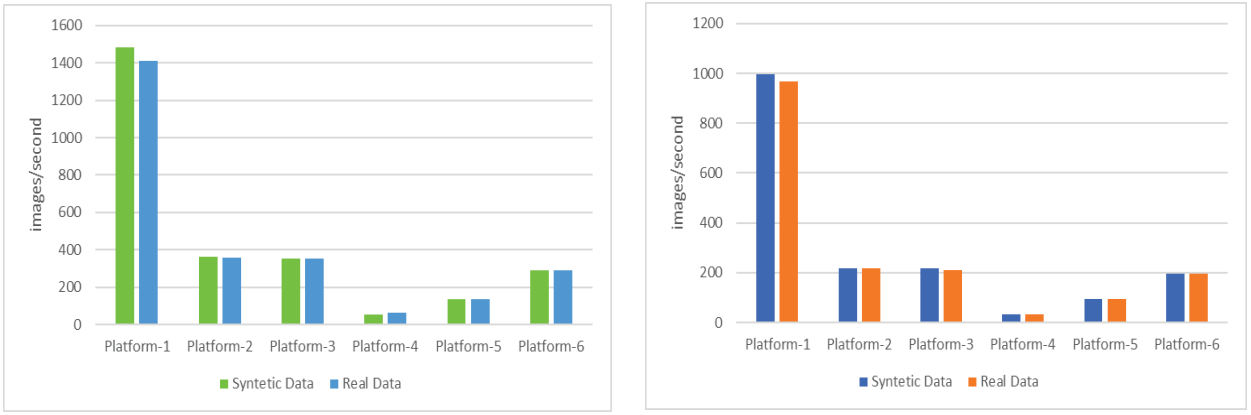


Fig. 3. Performance comparison when running (a) Inception v3 model and (b) Restnet-50 model using platforms described in Table 3. Similar to Fig.2, the higher the value (images s^{-1}) the better.

is quite similar to the commonly used cloud computing services: Google Compute Engine and Amazon EC2, each with eight GPU cards.

Knowing this, that this platform has a potential use case in future real-world applications especially in a country such as Indonesia since the access to the cloud computing services are expensive in long term scenario [16]. Arguably, this platform can easily outperform the dedicated Deep Learning computing engine such as Nvidia DGX-1 provided more similar cards were added to PakCarik, as shown by comparing Figure 2 and Figure 3 [17].

4. CONCLUSION AND SUGGESTIONS

After some experiments with PakCarik, we are convinced that PakCarik is ready to be deployed in real-world applications. In several scenarios as explained, PakCarik outperforms other similar but more expensive platforms available on the market. The authors measured its maximum computing performance at 326 Gflops. Furthermore, it can be inferred from the same experiment that PakCarik can be made even more powerful by adding more GPU cards.

5. ACKNOWLEDGEMENTS

This research project was supported partially by the Indonesian Ministry of Research and Technology/ National Research and Innovation Agency through the PDUPT Grant No. 007/SP2H/PDUPT/LPPM-UKP/IV/2021 and also by the Bureau of Research and Community Service as well as the Faculty of Industrial Technology at the Petra Christian University through several research funding: No. 01/PNLT/FTI/UKP/2018. The authors express sincere gratitude for the supports.

6. CONFLICT OF INTEREST

The authors declare no conflict of interest.

7. REFERENCES

1. A. Ng, Baidu's *Chief Scientist on Intersection of Supercomputing,* Machine Learning. [Online] from www.nextplatform.com/2016/04/01/baidus-chief-scientist-intersection-supercomputing-machine-learning/ (2016). [Accessed on July 27th 2019].
2. H. Nasser, Y. Hafeer and S. Ali. Towards software testing as a service for software as a service based on cloud computing model. *Proceedings of the Pakistan Academy of Sciences A. Physical and Computational Sciences* 55 (4): 1–8 (2018)
3. S.L. Graham., M. Snir, and C.A. Patterson. *Getting Up to Speed: The Future of Supercomputing.* National Academies Press (2005). DOI: 10.17226/11148
4. U. Khan and U. Naeem. Practices for clients in the adoption of hybrid cloud. *Proceedings of the Pakistan Academy of Sciences A Physical and Computational Sciences.* 54 (1): 13–32 (2017)
5. V.V. Kindratenko., J.J. Enos., G. Shi., M.T. Showerman., G.W. Arnold, J.E. Stone., J.C. Phillips, and W.M. Hwu, GPU clusters for high-performance computing. IEEE International Conference on Cluster Computing and Workshops (CLUSTER'09), (2009) pp. 1–8. DOI: 10.1109/CLUSTR.2009.5289128
6. *NVIDIA. NVIDIA DGX-1 With Tesla V100 System Architecture.* [Online] from http://images.nvidia.com/content/pdf/dgx1-v100-system-architecture-whitepaper.pdf (2018). [Accessed on July 27th 2019].
7. N.A. Gawande., J.A. Daily., C. Siegel., N.R. Tallent, and A. Vishnum, Scaling deep learning workloads: Nvidia dgx-1/pascal and intel knights landing. *Future Generation Computer Systems* 108:1162–1172 (2020). DOI:10.1016/j.future.2018.04.073
8. X. Glorot, and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), (Sardinia, Italy, 2010).
9. C. Berry., G. Hall., B. Matuszewski, and L.K. Shark. A comparison of architectures and evaluation of metrics for in-stream machine learning algorithms in industry 4.0 applications. 30th International Conference on Condition Monitoring and Diagnostic Engineering Management. (Preston and Grange-Over-Sands, UK, 2017).
10. L.Y. Joo., T.S. Yin., E. Xu., E. Thia., P.F. Chia., C.W.K. Kuah, and K.K. He, A feasibility study using interactive commercial off-the-shelf computer gaming in upper limb rehabilitation in patients after stroke. *Journal of Rehabilitation Medicine,* 42(5):437–441 (2010). DOI: 10.2340/16501977-0528
11. C. Szegedy., V. Vanhoucke., S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for

computer vision. The IEEE conference on computer vision and pattern recognition, pp. 2818–2826. (2016).

12. H. Kaiming., X. Zhang., R. Shaoqing, and J. Sun. Deep residual learning for image recognition. The IEEE conference on computer vision and pattern recognition, pp. 770–778. (2016).
13. Tensorflowx.org. Benchmarks. [Online] from https://www.tensorflow.org/guide/performance/benchmarks [Accessed on August 9th 2019].
14. A.Lee, C. Yau, M.B. Giles, A. Doucet and C.C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4): 769–789 (2010).
15. S. Shi, Q. Wang, P. Xu and X. Chu. Benchmarking state-of-the-art deep learning software tools. 7th International Conference on Cloud Computing and Big Data (CCBD), 2016, pp. 99–104. (2016). DOI: 10.1109/CCBD.2016.029.
16. R. Makhlouf,. Cloudy transaction costs: A dive into cloud computing economics. *Journal of Cloud Computing,* 9 (1):1–11. (2020). DOI: 10.1186/s13677-019-0149-4
17. O. Adjoua L. Lagardère, L.H. Jolly, A. Durocher, T. Very , I. Dupays, Z. Wang, T.J. Inizan , F. Célerse, P. Ren , J.W Ponder, J.P. Piquemal Tinker-HP: Accelerating molecular dynamics simulations of large complex systems with advanced point dipole polarizable force fields using GPUs and multi-GPU systems. *Journal of Chemical Theory and Computation,* 17 (4), 2034–2053 (2021). DOI: 10.1021/acs.jctc.0c01164