INDONESIA OpenInfra Days

02.11.2019 | Golden Tulip Legacy Hotel, Surabaya

PakCarik

GPU-Accelerated Platform for AI Research

by

Indar Sugiarto (indar@ieee.org)

Biznet GioCloud Biznet Mellanox TECHNOLOGIES BANK BRI OSF OpenStack Foundation

Contents

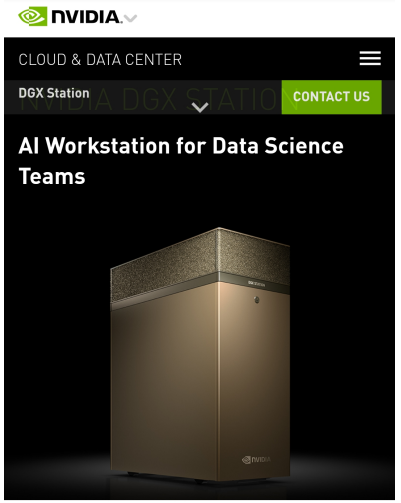
1. Introduction
   Motivation and Basic Ideas
2. Platform
   Hardware and Software
3. Performance
   What we've achieved
4. Application
   Our vision on applications
5. Conclusions
   Some thoughts

Biznet GioCloud Biznet Mellanox TECHNOLOGIES BANK BRI OSF OpenStack Foundation

Introduction

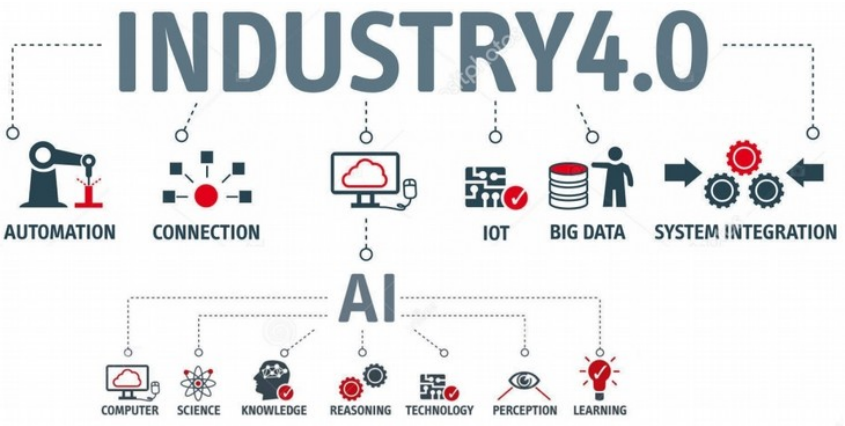
Motivation

1. Deep/Machine Learning is hungriest computation on data
2. Cloud computing is not cheap for sensitive/heavy load
3. Dedicated machines from vendors (e.g., NVIDIA DGX series) are expensive
4. We need a fully customizable and easily configurable platform



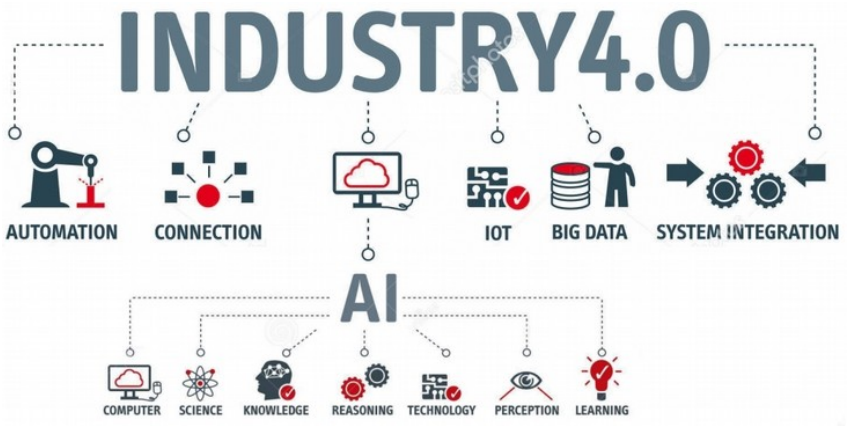
Basic Ideas :: AI Era

We need **AI** these days...

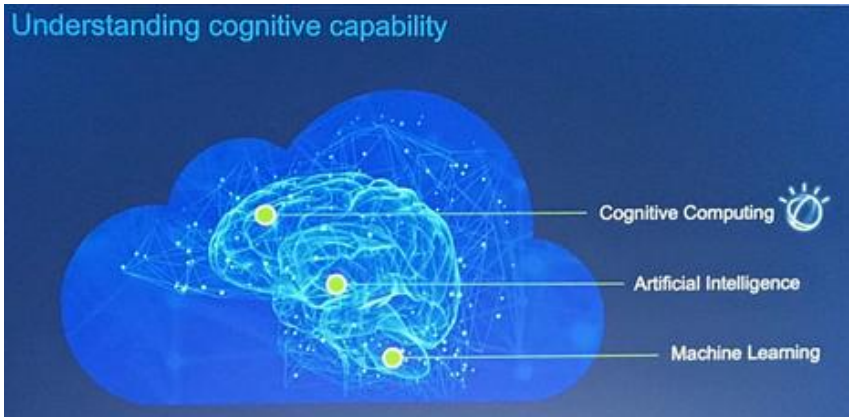


Basic Ideas :: AI Era

We need **AI** these days...

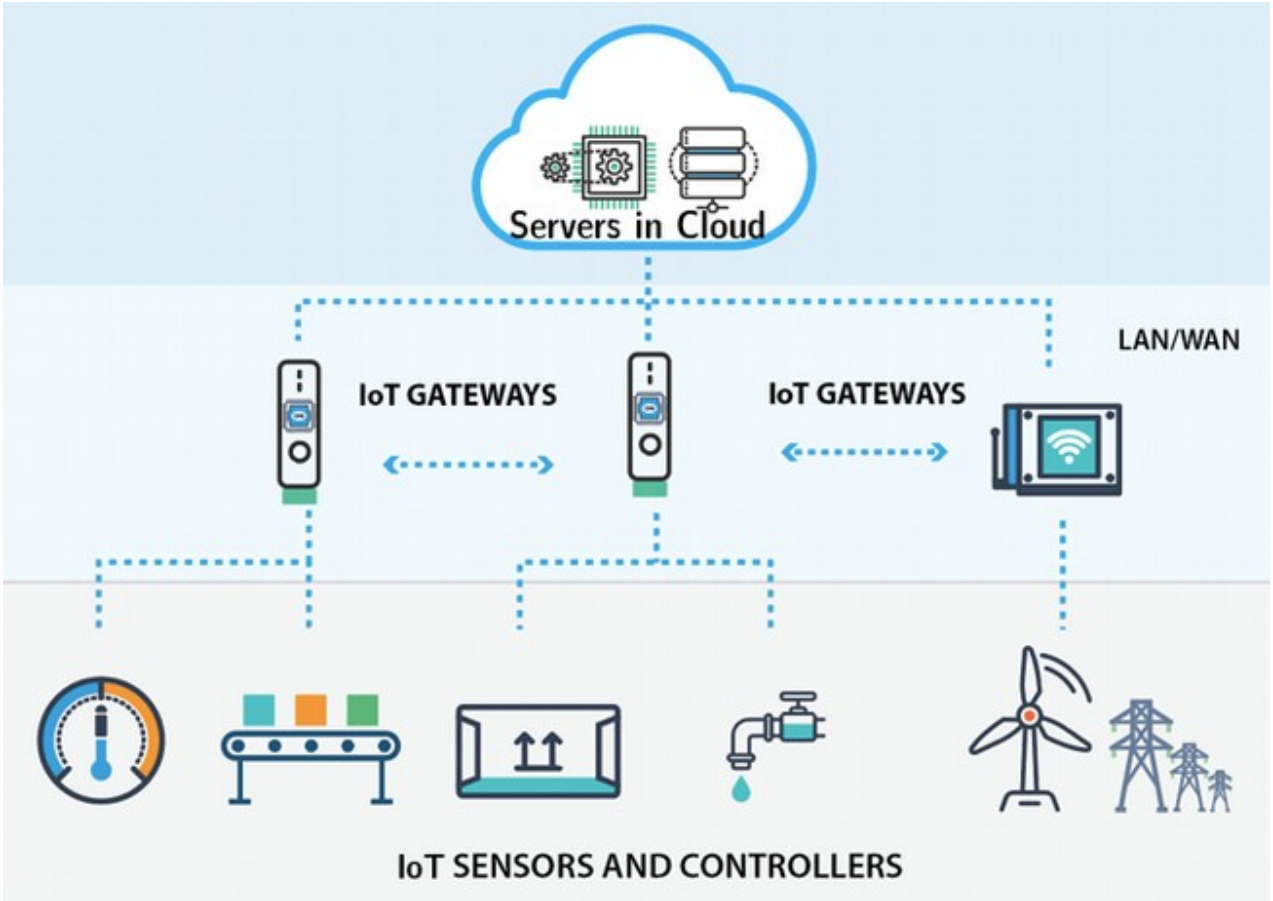


Not just AI, but **Cognitive Computing**...



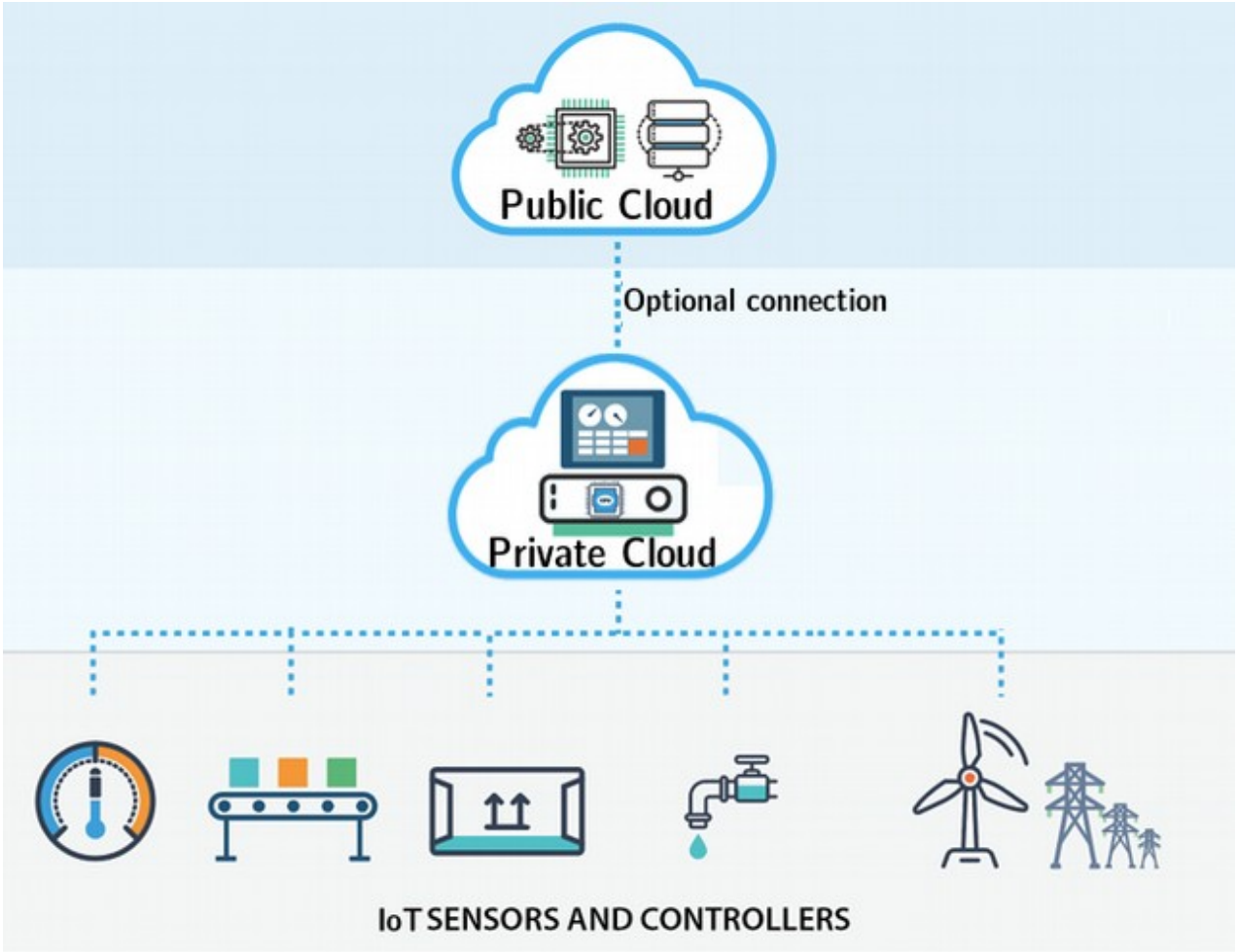
Basic Ideas :: Cloud Computing

Usual case in IoE world...



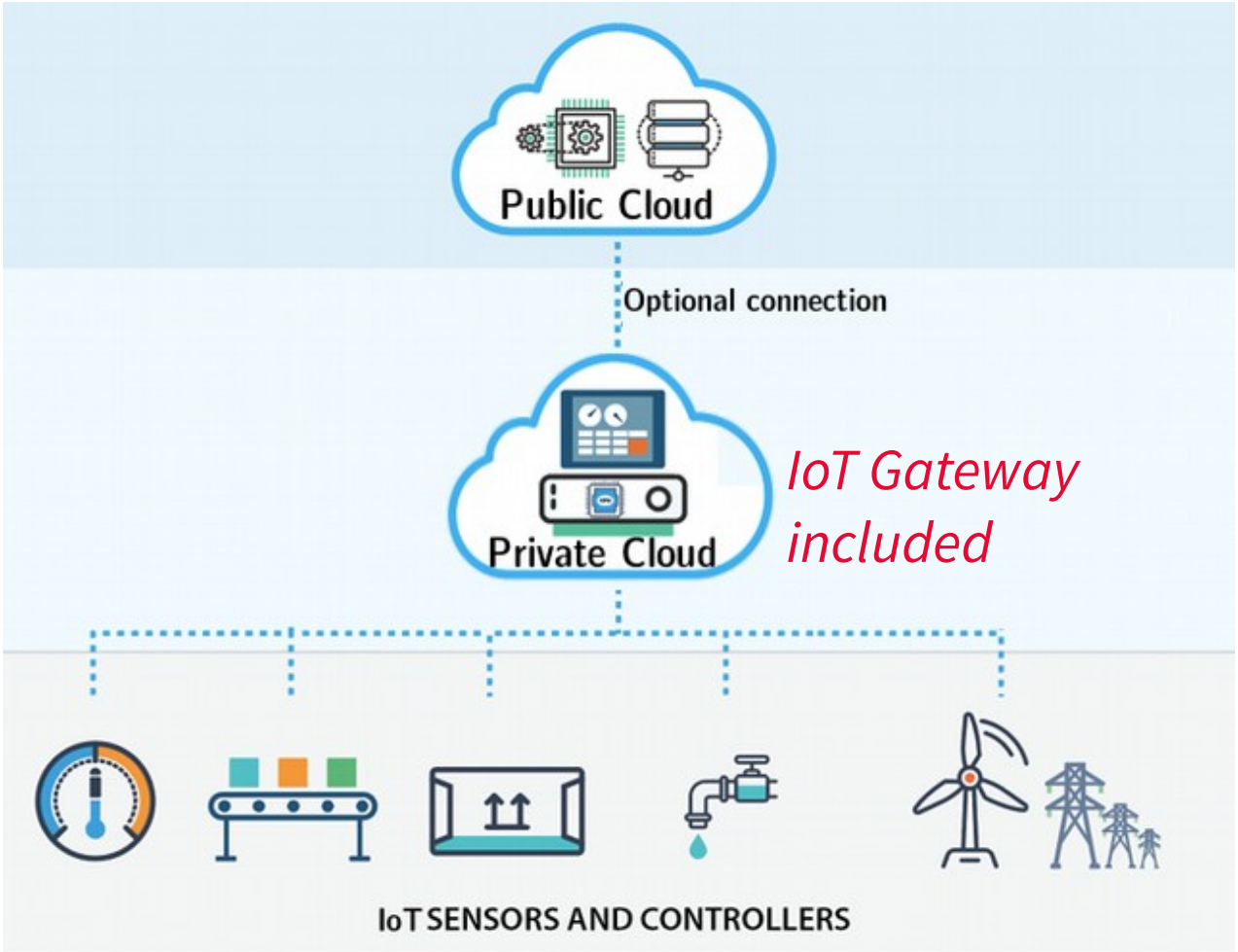
Basic Ideas :: Cloud Computing

Optimized IoE world...



Basic Ideas :: Cloud Computing

Optimized IoE world...



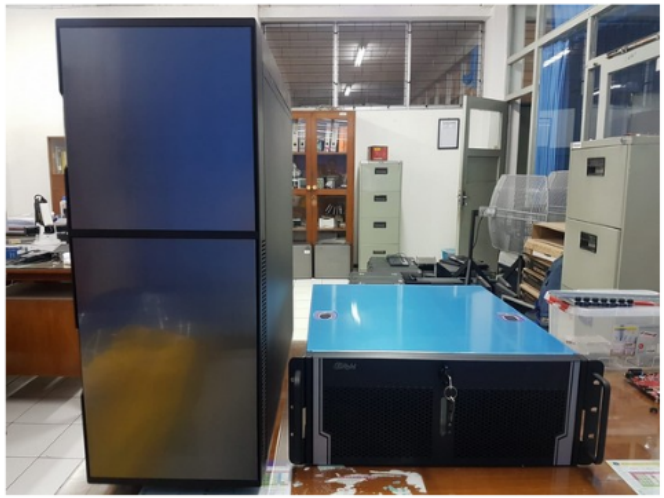
Platform

Introduction

- **PakCarik** is an Indonesian acronym for **Platform Komputasi Cerdas Ramah Industri Kreatif** (”Creative Industry friendly Intelligence Computing Platform“)
- Aims: to provide a complete development and production environment for AI-based projects, especially to those that rely on machine and deep learning paradigms.
- **PakCarik** was built using commercial off-the-shelf hardware and tested on several application scenarios.

Biznet GioCloud Biznet Mellanox TECHNOLOGIES BANK BRI OSF OpenStack Foundation

Prototypes



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	1000 Watt	1200 Watt
Form factor	ATX-Tower	Rackmount 4U

Libraries

Current configuration:

- OS'es: Ubuntu Server 18.04 with Docker
- Compiler and scripting: GNU C/C++, Fortran, CUDA, OpenCL, OpenMP, MPI, Python, R, Octave
- Libraries for AI and machine learning: Caffe, TensorFlow, Theano, Keras, PyTorch, Scikit-Learn
- IoT and Bigdata Platforms: RabbitMQ, HDFS / Spark+Kafka, Cassandra, Thingsboard.io, etc.

Users can install/update packages automatically from our repository: `http://pakcarik.petra.ac.id`

Difficulties

Our headache was due to:

- Tensorflow-gpu: Ubuntu is preferrable, Debian is doable, others are complicated
- CUDA Toolkit 10.x is nice, but was not usable (until recently)
- PC1 (with i9 and TBB) runs HPL ”smoother“ than PC2 (Threadripper 2990WX)
- Competing/conflicting IoT-gateways: RabbitMQ (for AMQP) vs Thingsboard (for MQTT) vs Hadoop (Kafka)
- PC2 didn’t boot when in MAAS, but PC1 did
- and those intermittent stuffs...

Difficulties

Our headache was due to:

- Tensorflow-gpu: Ubuntu is preferrable, Debian is doable, others are complicated
- CUDA Toolkit 10.x is nice, but was not usable (until recently)
- PC1 (with i9 and TBB) runs HPL ”smoother“ than PC2 (Threadripper 2990WX)
- Competing/conflicting IoT-gateways: RabbitMQ (for AMQP) vs Thingsboard (for MQTT) vs Hadoop (Kafka)
- PC2 didn’t boot when in MAAS, but PC1 did
- and those intermittent stuffs...

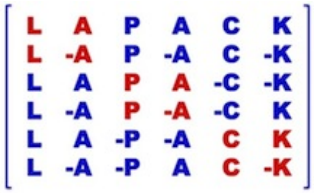
Too many ideas (from our researchers). Any solution?

Performance

How did we evaluate?

In order to evaluate the performance of **PakCarik**, we devised several evaluation scenarios:

1. Profiling computation performance based on a linear algebra problem solver (we used high performance Lapack - **HPL**)
2. Distributed computing using message passing interface (**MPI**) protocol
3. Intensive computing for image processing using Deep Learning frameworks (developed using TensorFlow – **TF**)



Result :: No Accelerator

Testing on CPU only (without GPU acceleration)

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

Note: PC1 run at 3.3GHz, PC2 run at 3.0GHz

Observation?

Result :: No Accelerator

Testing on CPU only (without GPU acceleration)

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

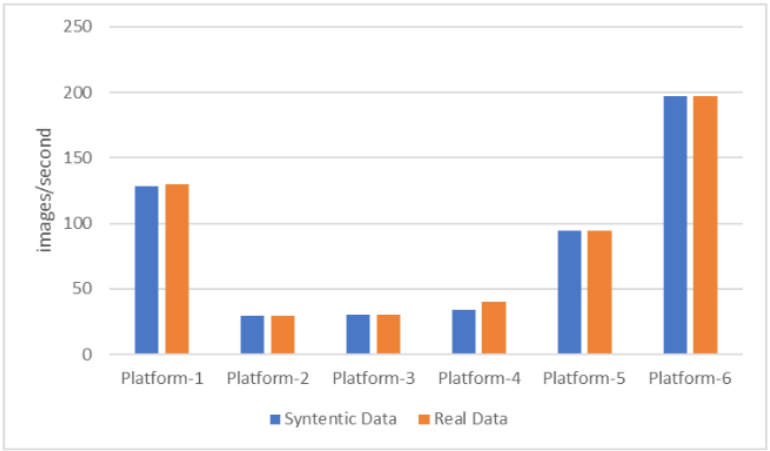
Note: PC1 run at 3.3GHz, PC2 run at 3.0GHz

Observation? not much different!

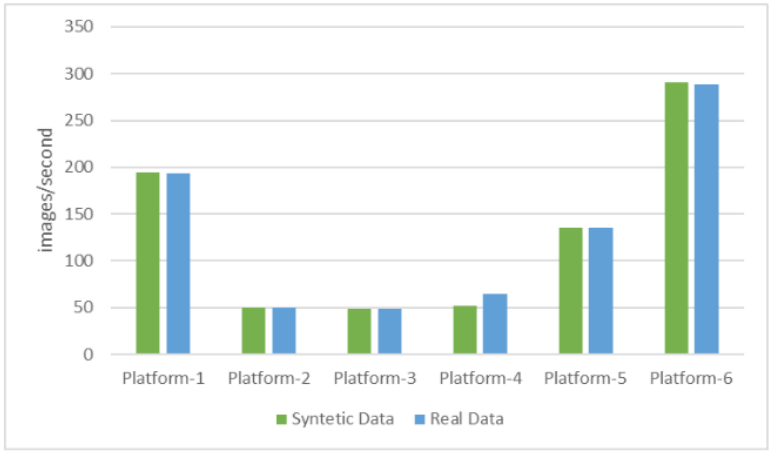
Result :: With Acceleration

Testing with GPU acceleration for Image Processing using Deep Learning framework

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)
Platform-6	PC2 (PakCarik prototype 2)



Performance when running Inception v3



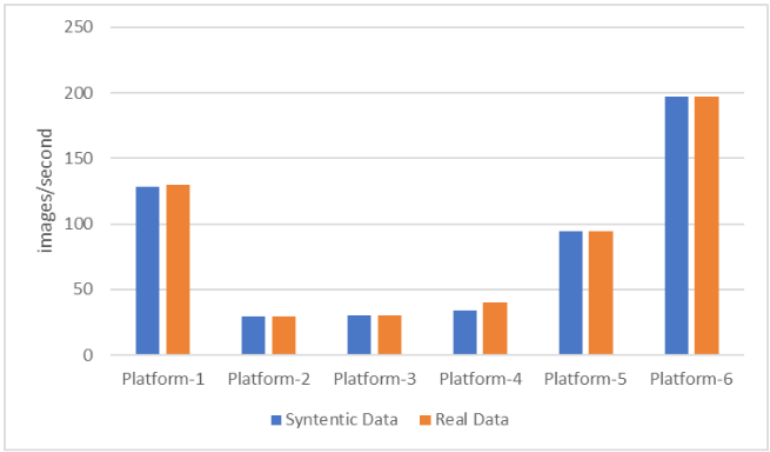
Performance when running Restnet-50

Result :: With Acceleration

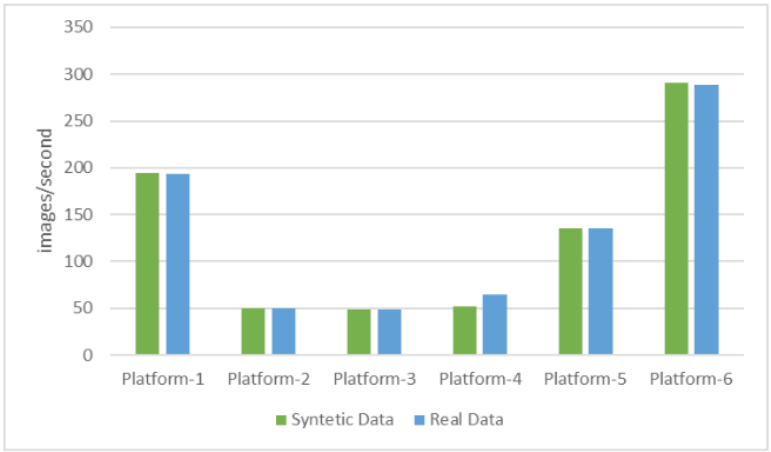
Testing with GPU acceleration for Image Processing using Deep Learning framework

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)
Platform-6	PC2 (PakCarik prototype 2)

1 GPU



Performance when running Inception v3

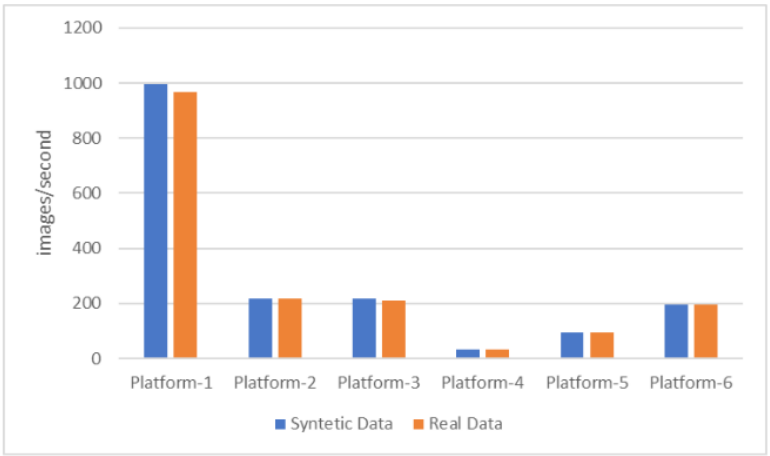


Performance when running Restnet-50

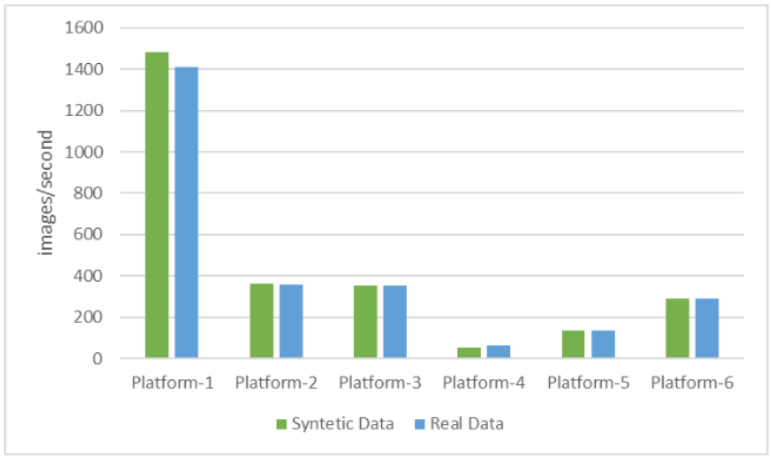
Result :: With Acceleration

Testing with GPU acceleration for Image Processing using Deep Learning framework

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)
Platform-6	PC2 (PakCarik prototype 2)



Performance when running Inception v3



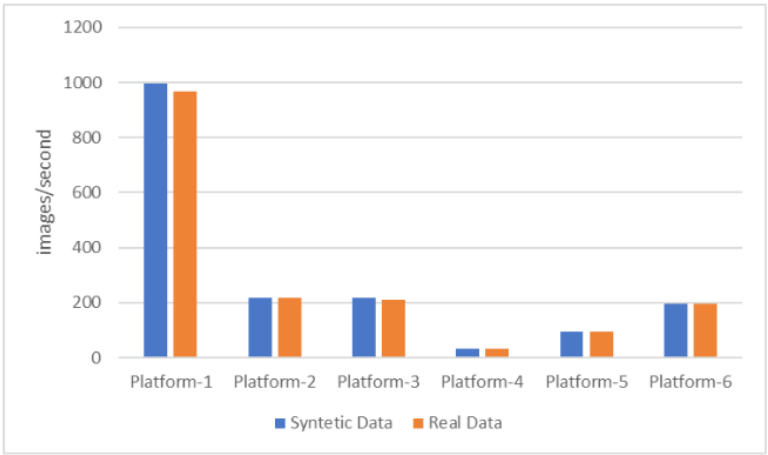
Performance when running Restnet-50

Result :: With Acceleration

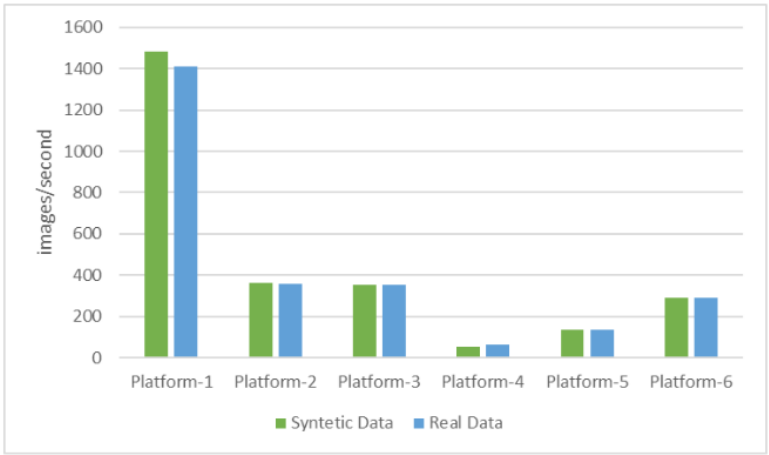
Testing with GPU acceleration for Image Processing using Deep Learning framework

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)
Platform-6	PC2 (PakCarik prototype 2)

8 GPUs



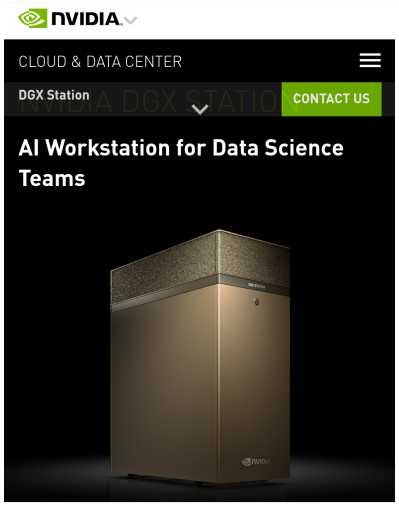
Performance when running Inception v3



Performance when running Restnet-50

Result :: What can we learn?

1. Obviously, NVIDIA DGX-1 with **8 GPU** cards outperforms all other platforms
2. The performance of PC2 with only one GPU card is quite similar to Google Compute Engine and Amazon EC2, (**both with 8 GPU cards**)
3. Our platform can easily outperform the dedicated Deep Learning computing engine such as NVIDIA DGX-1 provided more cards were added to **PakCarik**



amazon web services EC2

Google Compute Engine

Biznet GioCloud

Biznet

Mellanox TECHNOLOGIES

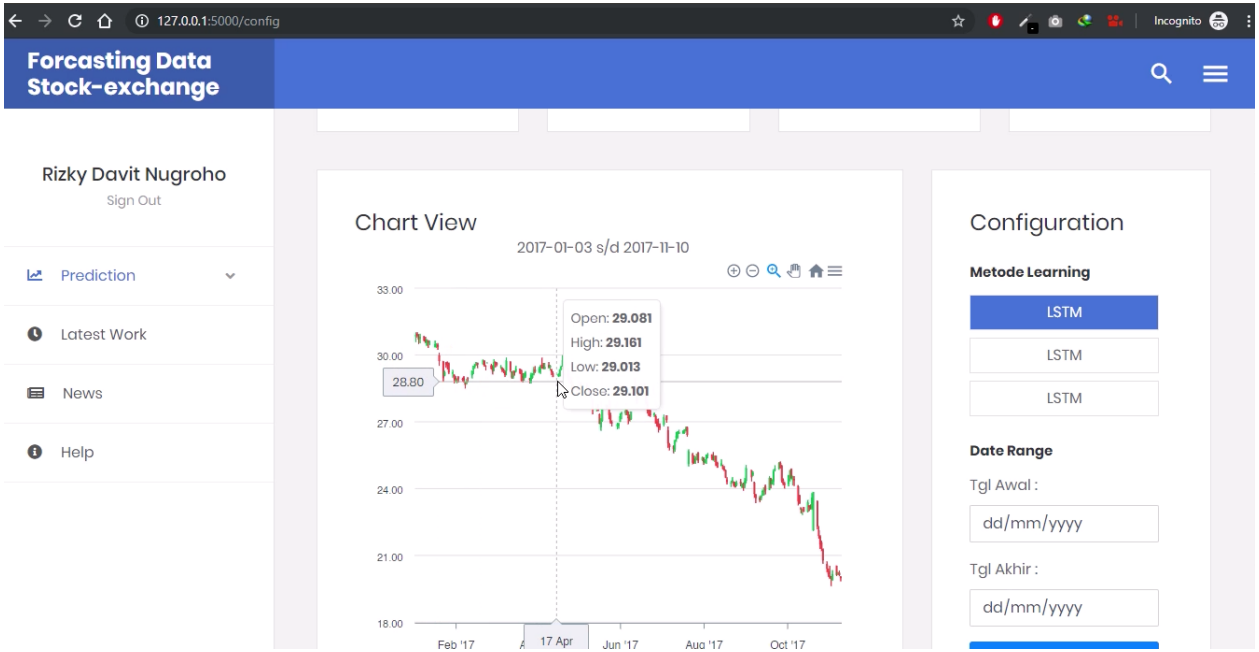
BANK BRI

OSF OpenStack Foundation

Applications

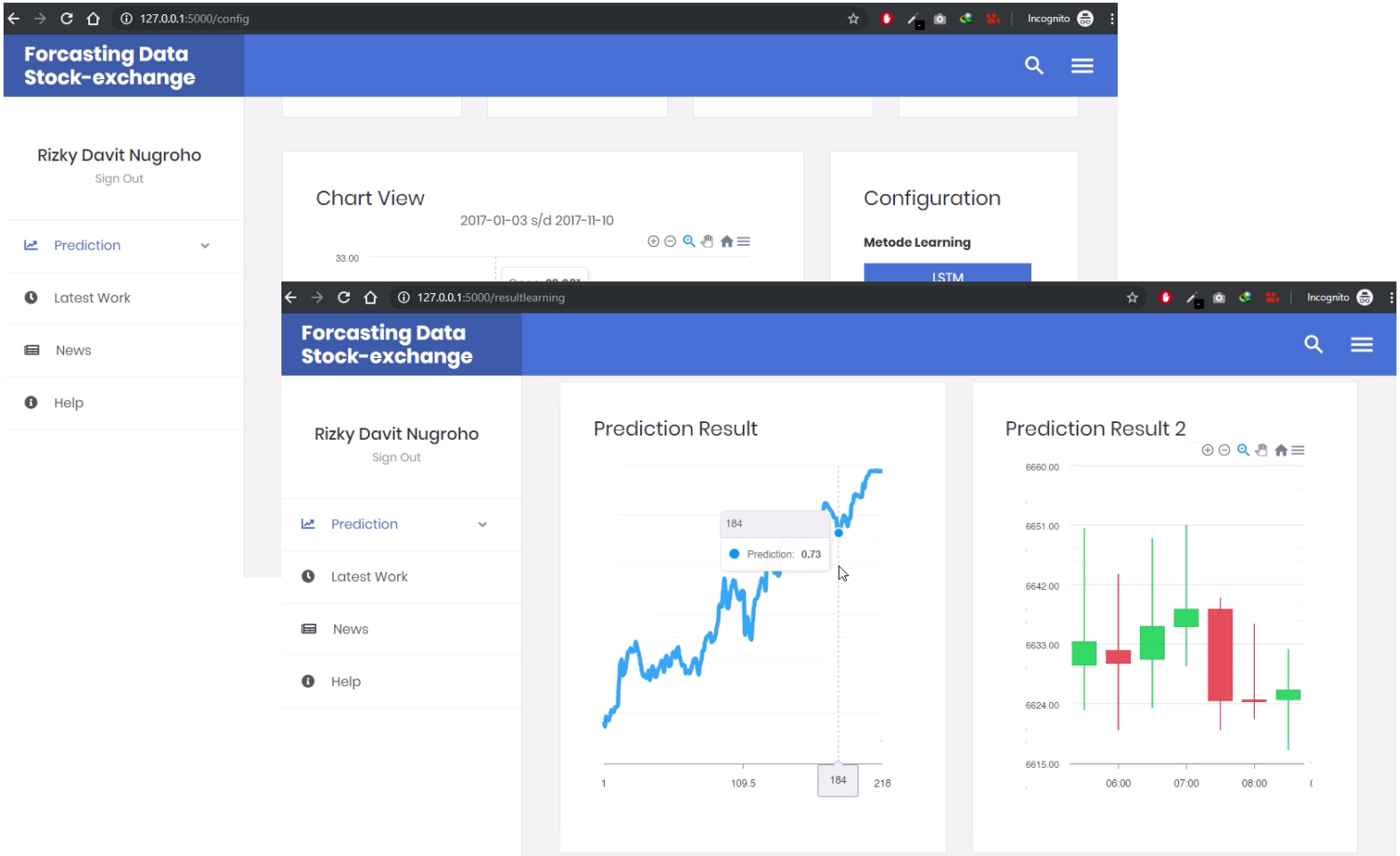
Time Series Forecasting

Stock Market Prediction



Time Series Forecasting

Stock Market Prediction



UPAIPRO.COM

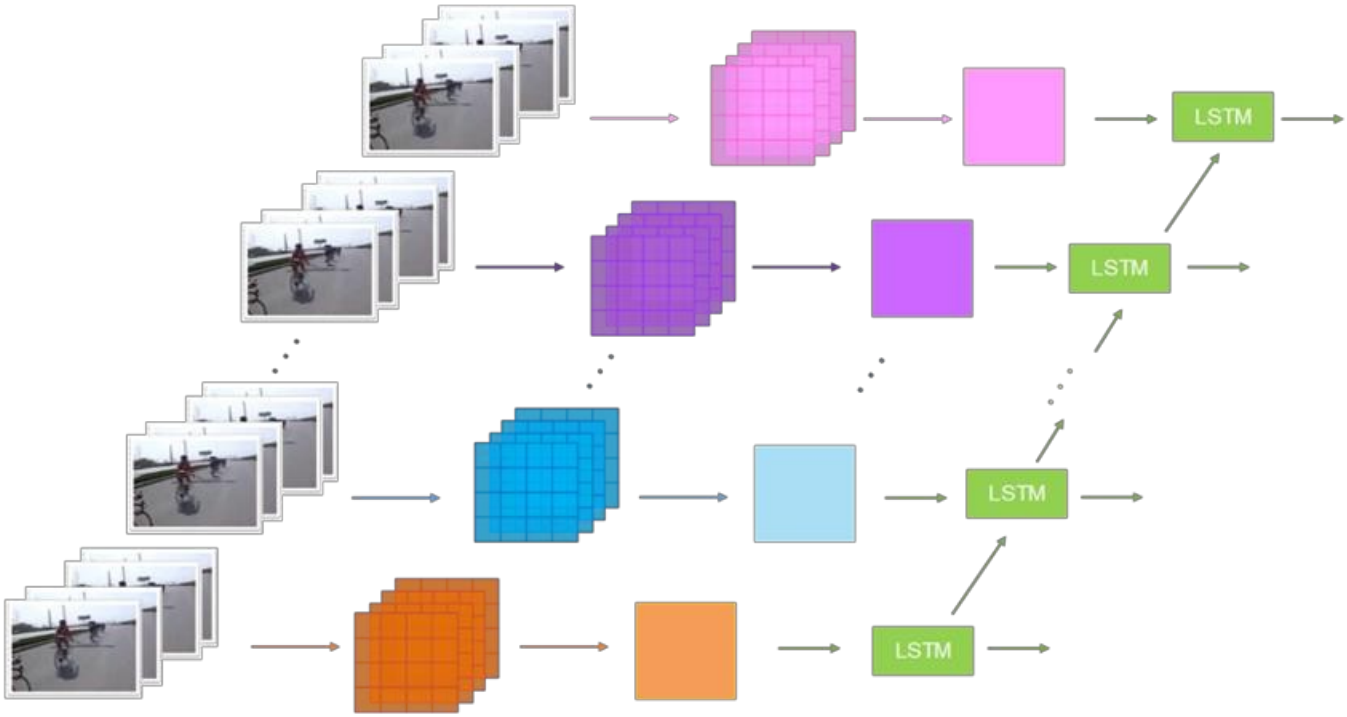

Unlimited Profit Professional
Hello , Indar Sugiarto
Unlimited Profit Professional
AI-based Business Analytic Platform
UPAIPRO
Home
Service
Contact Us
Logout
RECOMMENDED
Fokus 10 Saham Pilihan UPAIPRO untuk Bulan Oktober 2019:
ACES – ANTM – BBCA – ICBP – KLBF – MAPI – MNCN – MTDL – SMRA – WOOD
Keterangan singkat dari saham-saham yang direkomendasikan oleh UPAIPRO bisa dilihat di sini!
Silahkan pilih dari daftar saham berikut untuk melihat detil hasil prediksinya.
ACE The helpful place.
adhi
ADIRA FINANCE
antam
ASTRA international
ASTRA Otoparts
BCA
BANK BUKOPIN
BNI

Current Work :: Task

Emergency Event Detection Based on Crowd Behaviour Analysis

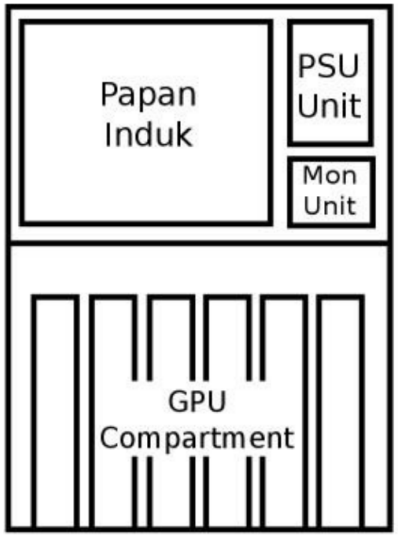
Current Work :: Algorithm

Visual attention guided 3D CNN



Current Work :: GPU cluster

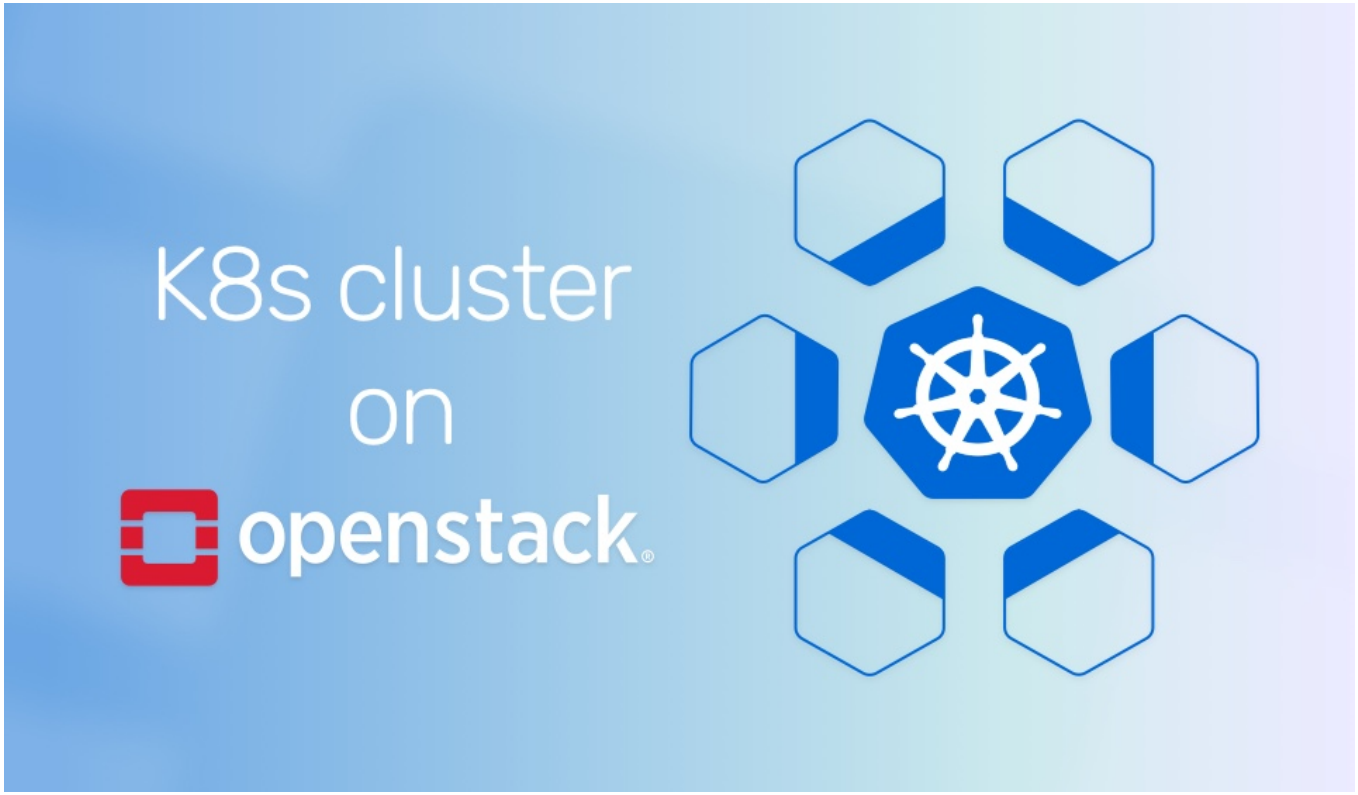
What we need



Gambar 11. Rencana ekpansi **Pak Carik** menjadi GPU-farm yang menjadi bagian dari *cluster supercomputer*.

Current Work :: Private Cloud

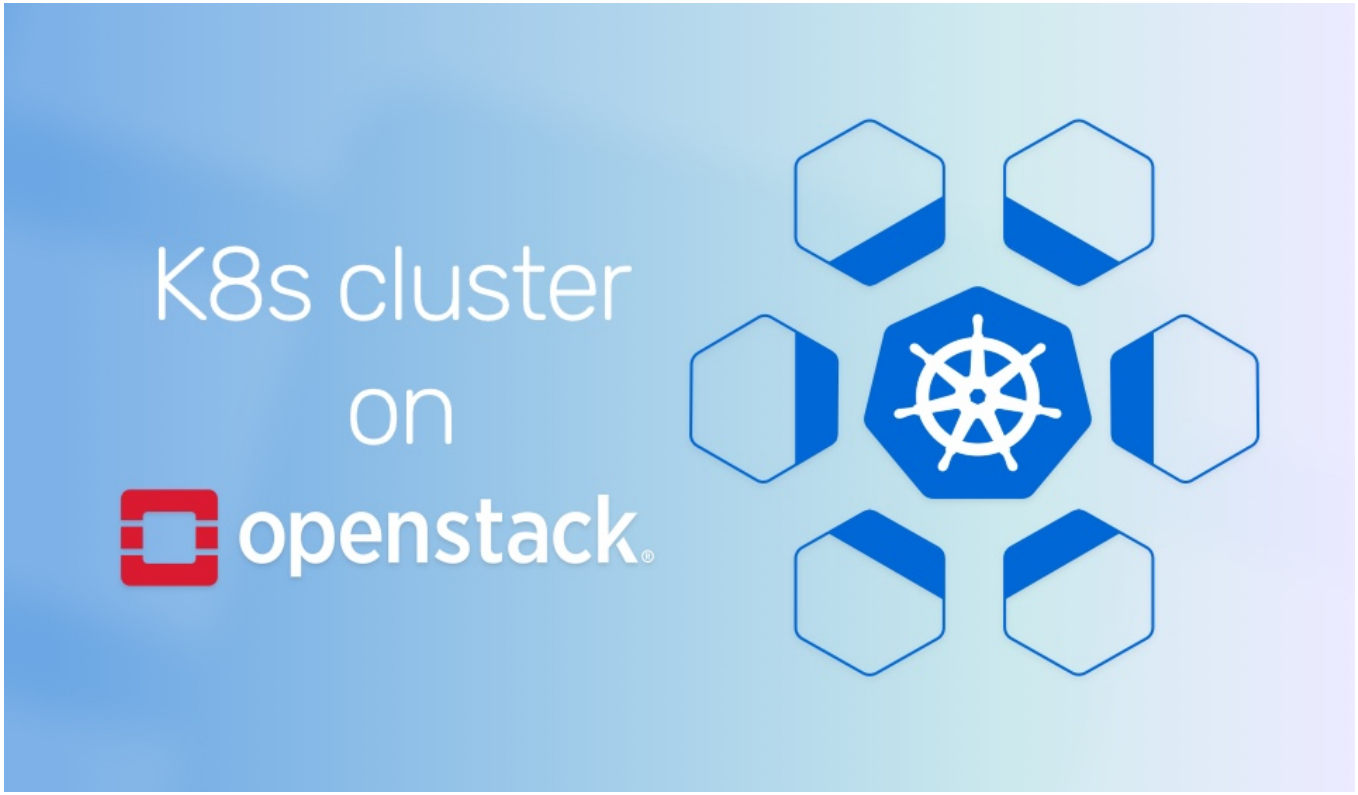
Kubernets on Openstack



Biznet GioCloud Biznet Mellanox TECHNOLOGIES BANK BRI OSF OpenStack Foundation

Current Work :: Private Cloud

Kubernets on Openstack



How scalable is it? Maintainability?

Biznet GioCloud Biznet Mellanox TECHNOLOGIES BANK BRI OSF OpenStack Foundation

Our Vision


Society 5.0
• Create jobs
• Reduce polarization
• Improve quality of life
Engines of 4th industrial revolution
Big data
Artificial intelligence
AIoT
Internet of Things
Cloud

Conclusions

Closing Thoughts

- AI with Deep/Machine Learning is ”hot“ domain but it is the hungriest system on computation resources

Closing Thoughts

- AI with Deep/Machine Learning is ”hot“ domain but it is the hungriest system on computation resources
- Building a fully customizable and easily configurable platform with COTS components is doable

Closing Thoughts

- AI with Deep/Machine Learning is ”hot“ domain but it is the hungriest system on computation resources
- Building a fully customizable and easily configurable platform with COTS components is doable
- More researches/experiments on the platform, using either HPC- or HTC approach, is needed to fully harness the power of the Private Cloud

Closing Thoughts

- AI with Deep/Machine Learning is ”hot“ domain but it is the hungriest system on computation resources
- Building a fully customizable and easily configurable platform with COTS components is doable
- More researches/experiments on the platform, using either HPC- or HTC approach, is needed to fully harness the power of the Private Cloud
- We invite you for **collaboration!**

Closing Thoughts

- AI with Deep/Machine Learning is ”hot“ domain but it is the hungriest system on computation resources
- Building a fully customizable and easily configurable platform with COTS components is doable
- More researches/experiments on the platform, using either HPC- or HTC approach, is needed to fully harness the power of the Private Cloud
- We invite you for **collaboration!**

Thank you!

Our Team

Dr.Ing., Indar Sugiarto (Dept. Electrical Engineering, Petra Christian University, Indonesia)
Felix Pasila, PhD. (Dept. Electrical Engineering, Petra Christian University, Indonesia)
Resmana Lim, M.Sc. (Dept. Electrical Engineering, Petra Christian University, Indonesia)
Agustinus Noertjahyana, M.Sc. (Dept. Informatics, Petra Christian University, Indonesia)
Henry Palit, PhD. (Dept. Informatics, Petra Christian University, Indonesia)
I Gede Widyadana, PhD. (Dept. Industrial Engineering, Petra Christian University, Indonesia)
Dr.Ing., Surya Hermawan (Dept. Civil Engineering, Petra Christian University, Indonesia)
Agustinus Bimo Gumelar, M.Sc. (Dept. Information System, Narotama University, Indonesia)