SUMMIT
ONLINE

대용량 한글 자연어처리 모델의 클라우드 분산 학습 및 배포 사례

전희원
연구원
SK Telecom

김무현
시니어 데이터 사이언티스트
AWS Korea

강지양
시니어 딥러닝 아키텍트
AWS Korea

aws SUMMIT ONLINE

Korean GPT-2 (KoGPT2)

aws SUMMIT ONLINE

GPT(Generative Pre-Training)2 - 1

- Language Model based Transformer
  - Language Model
    - P(아버지가 방에 들어가신다.) > P(아버지 가방에 들어가신다.)
  - Unsupervised pre-training

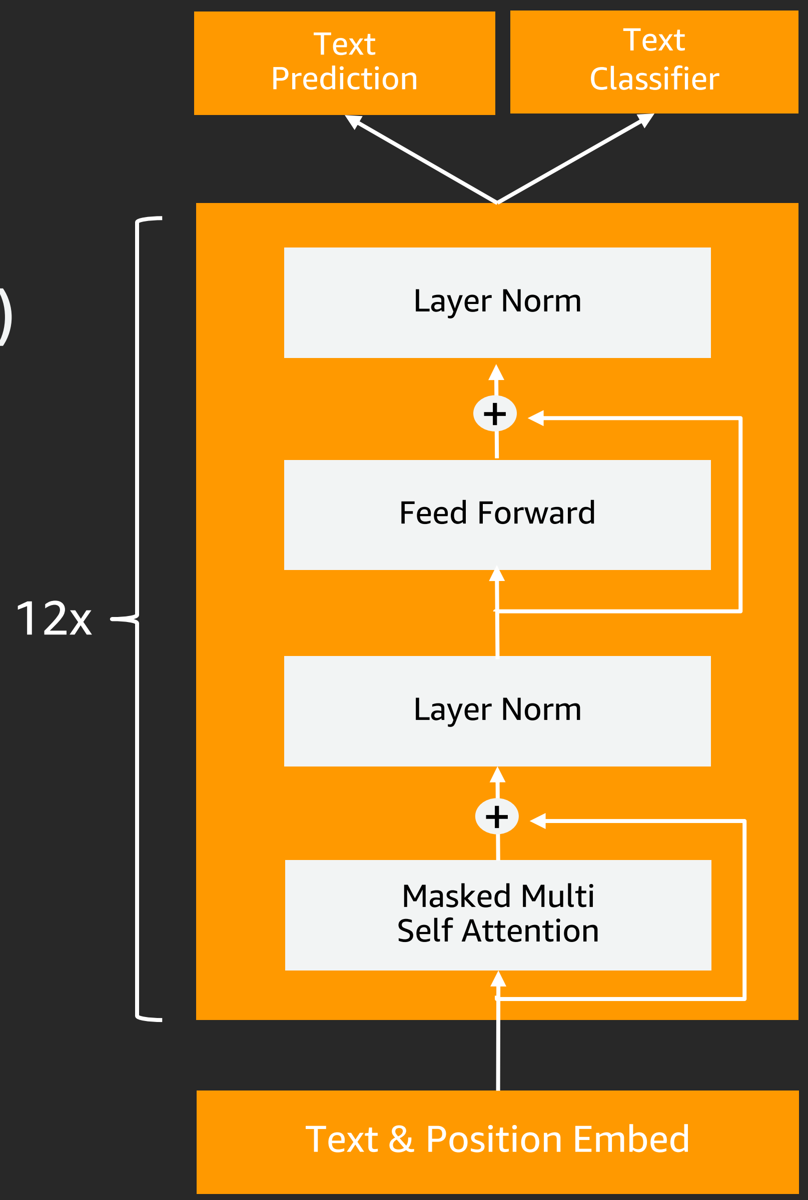
$$L_1(u) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

  - Transformer decoder

$$h_0 = UW_e + W_p$$

$$h_l = transformer\_block(h_{l-1}) \forall i \in [1, n]$$

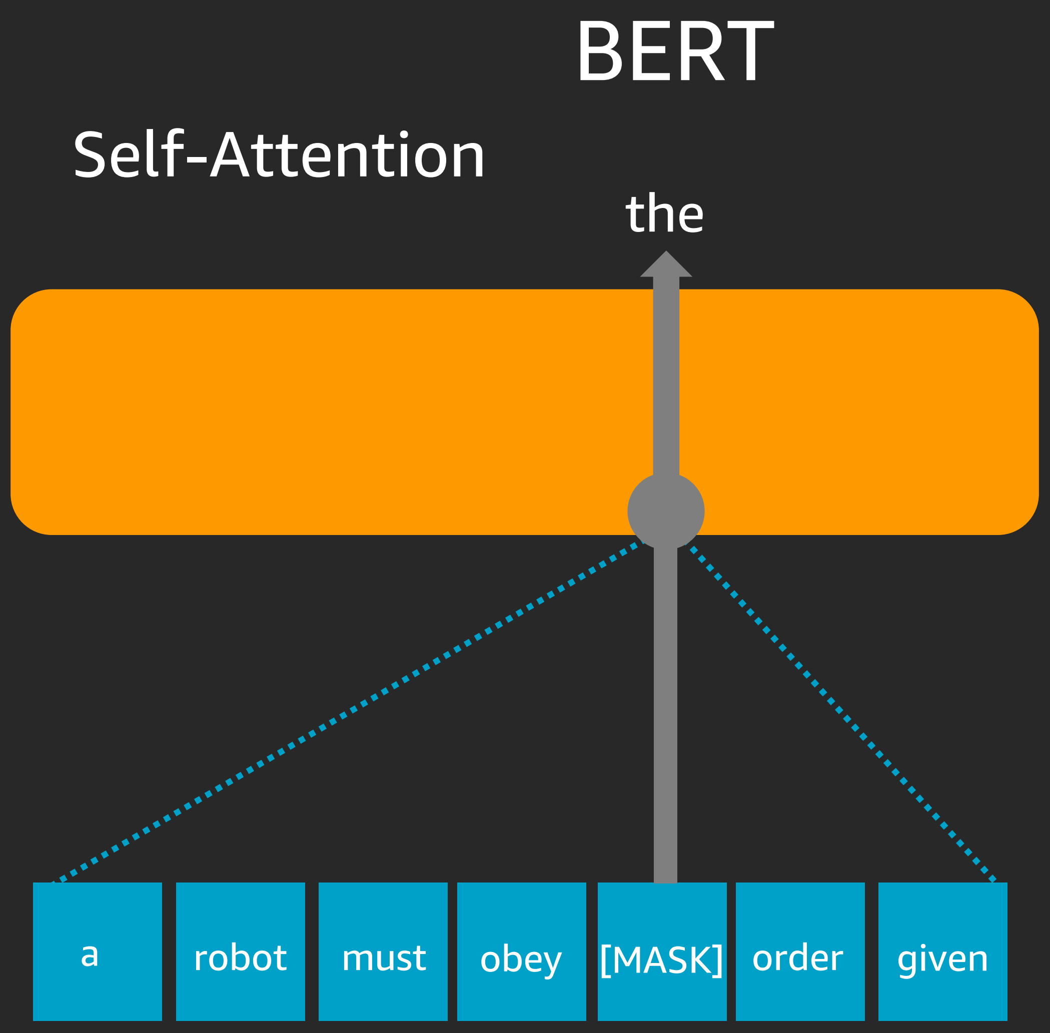
$$P(u) = softmax(h_n W_e^T)$$

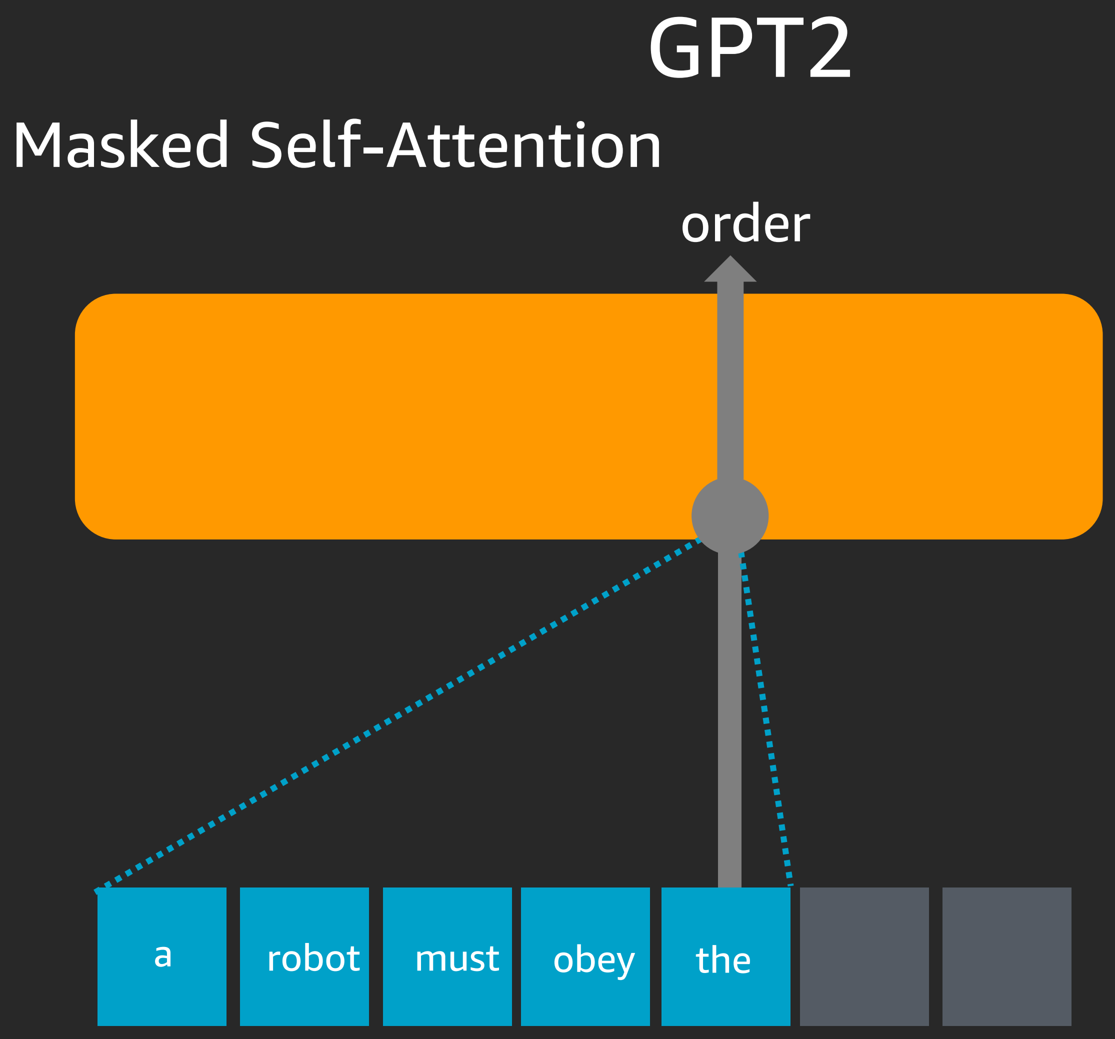


https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

GPT(Generative Pre-Training)2 - 2

> **A robot must obey the orders** given it by human beings except where such orders would conflict with the First Law.





http://jalammar.github.io/illustrated-gpt2/

KoGPT2 Training - Data

Corpus

Data	# of Sentences	# of Words
Korean Wiki	5M	54M
Korean News	120M	1.6B
Other corpus	9.4M, 18M	88M, 82M

- 20GB raw text (5 times larger than KoBERT)

Tokenizer

- Trained with 25M sentences(wiki + news)
- BPE(Byte Pair Encoding)
- 50,000 vocab

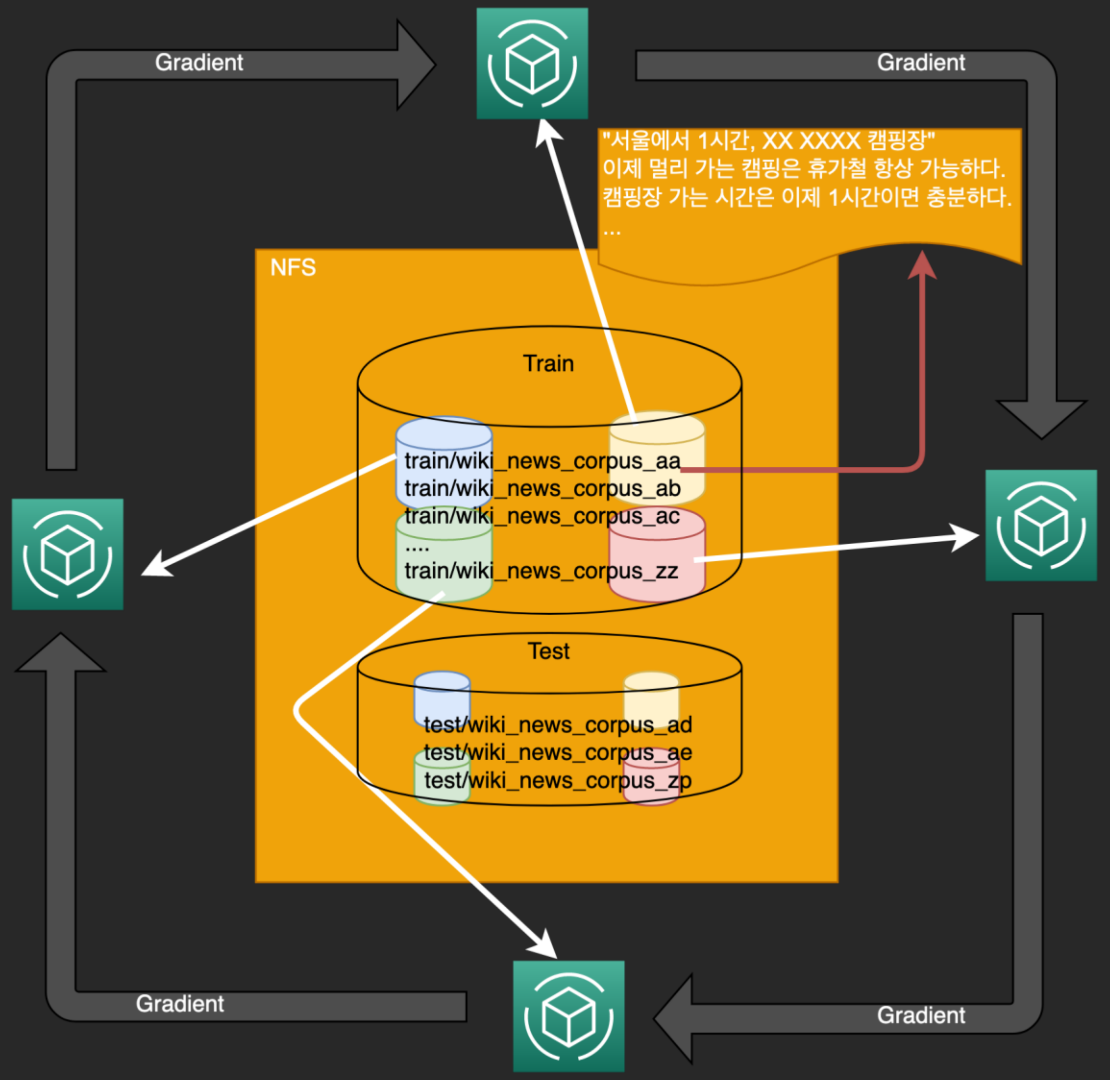
```
$ head dataset/wiki_bert_dataset_201901.txt
지미 카터
제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~ )는 민주당 출신 미국 39번째 대통령 (1977년 ~ 1981년)이다.
지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다.
조지아 공과대학교를 졸업하였다.
그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다.
1953년 미국 해군 대위로 예편하였고 이후 땅콩·면화 등을 가꿔 많은 돈을 벌었다.
그의 별명이 "땅콩 농부" (Peanut Farmer)로 알려졌다.
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1
대통령이 되기 전 조지아주 상원의원을 두번 연임했으며, 1971년부터 1975년까지 조지아 지사로 근무했다.
조지아 주지사로 지내면서, 미국에 사는 흑인 등용법을 내세웠다.
```

```
> tokenizer('news_wiki_2019_large_vocab.model', 'dataset/wiki_bert_dataset_201901.txt', 'dataset/wiki_bert
```

```
$ head  dataset/wiki_bert_dataset_201901_large_vocab_tok.txt
_지미 _카터
_제임스 _얼 _" 지 미 " _카터 _주니어 ( , _19 24 년 _10 월 _1 일 _ ~ _ ) 는 _민주당 _출신 _미국 _39 번째 _대통령 _ (19 77
_지미 _카터 는 _조지아주 _섬 터 _카운티 _플레 인 스 _마을에서 _태어났다 .
_조지 아 _공 과 대학교 를 _졸업 하였다 .
_그 _후 _해군 에 _들어가 _전 함 · 원자력 · 잠 수 함 의 _승무원 으로 _일 하였다 .
_1953 년 _미국 _해군 _대 위로 _예 편 하였고 _이후 _땅 콩 · 면 화 _등을 _가꿔 _많은 _돈을 _벌었다 .
_그의 _별명 이 _" 땅 콩 _농부 " _( P ean ut _F ar mer ) 로 _알려졌다 .
_1962 년 _조지 아 _주 _상원 _의원 _선거에서 _낙선 하나 _그 _선거 가 _부정선거 _ 였 음을 _입증 하게 _되어 _당선되 고 , _1966 년 _
_대통령이 _되기 _전 _조지아주 _상원의원 을 _두번 _연임 했으며 , _1971 년부터 _1975 년까지 _조지 아 _지사 로 _근무 했다 .
_조지 아 _주지사 로 _지내면서 , _미국에 _사는 _흑인 _등 용 법을 _내세웠다 .
```

KoGPT2 Training - Distributed training

- 'Single Reducer' to 'Ring Reducer'
  - Linear scale training performance with Horovod
- Instant corpus pre-processing
  - No need to load all data on memory.
- Syncing training chunks on every epoch.
  - No need to stop training to add more data.
- Fused GELU (with GluonNLP team)
  - About 15% performance improved



KoGPT2 Training - Performances

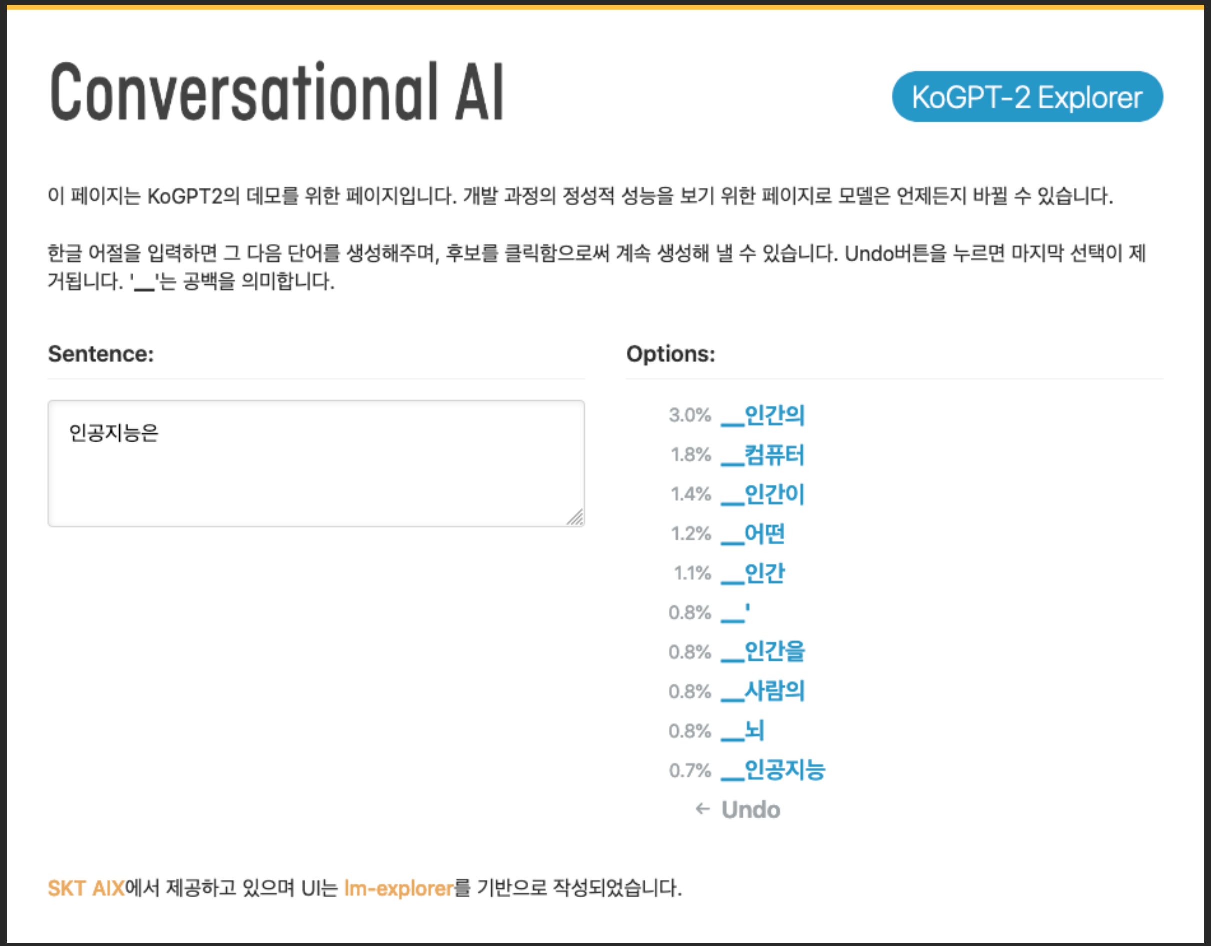
Sentiment Analysis (Naver movie review data)

Model	Test Accuracy
BERT base multilingual cased	0.875
KoBERT	0.901
KoGPT2	0.899

Paraphrase Detection

Model	Test Accuracy
KoBERT	0.912
KoGPT2	0.943

KoGPT2 Demo


Conversational AI
KoGPT-2 Explorer
이 페이지는 KoGPT2의 데모를 위한 페이지입니다. 개발 과정의 정성적 성능을 보기 위한 페이지로 모델은 언제든지 바뀔 수 있습니다.
한글 어절을 입력하면 그 다음 단어를 생성해주며, 후보를 클릭함으로써 계속 생성해 낼 수 있습니다. Undo버튼을 누르면 마지막 선택이 제거됩니다. '__'는 공백을 의미합니다.
Sentence:
인공지능은
Options:
3.0% __인간의
1.8% __컴퓨터
1.4% __인간이
1.2% __어떤
1.1% __인간
0.8% __'
0.8% __인간을
0.8% __사람의
0.8% __뇌
0.7% __인공지능
← Undo
SKT AIX에서 제공하고 있으며 UI는 lm-explorer를 기반으로 작성되었습니다.

Conversational AI

KoGPT-2 Explorer

이 페이지는 KoGPT2의 데모를 위한 페이지입니다. 개발 과정의 정성적 성능을 보기 위한 페이지로 모델은 언제든지 바뀔 수 있습니다.

한글 어절을 입력하면 그 다음 단어를 생성해주며, 후보를 클릭함으로써 계속 생성해 낼 수 있습니다. Undo버튼을 누르면 마지막 선택이 제거됩니다. '__'는 공백을 의미합니다.

Sentence:

여러분 2019년 수고하셨습니다. 올해에는

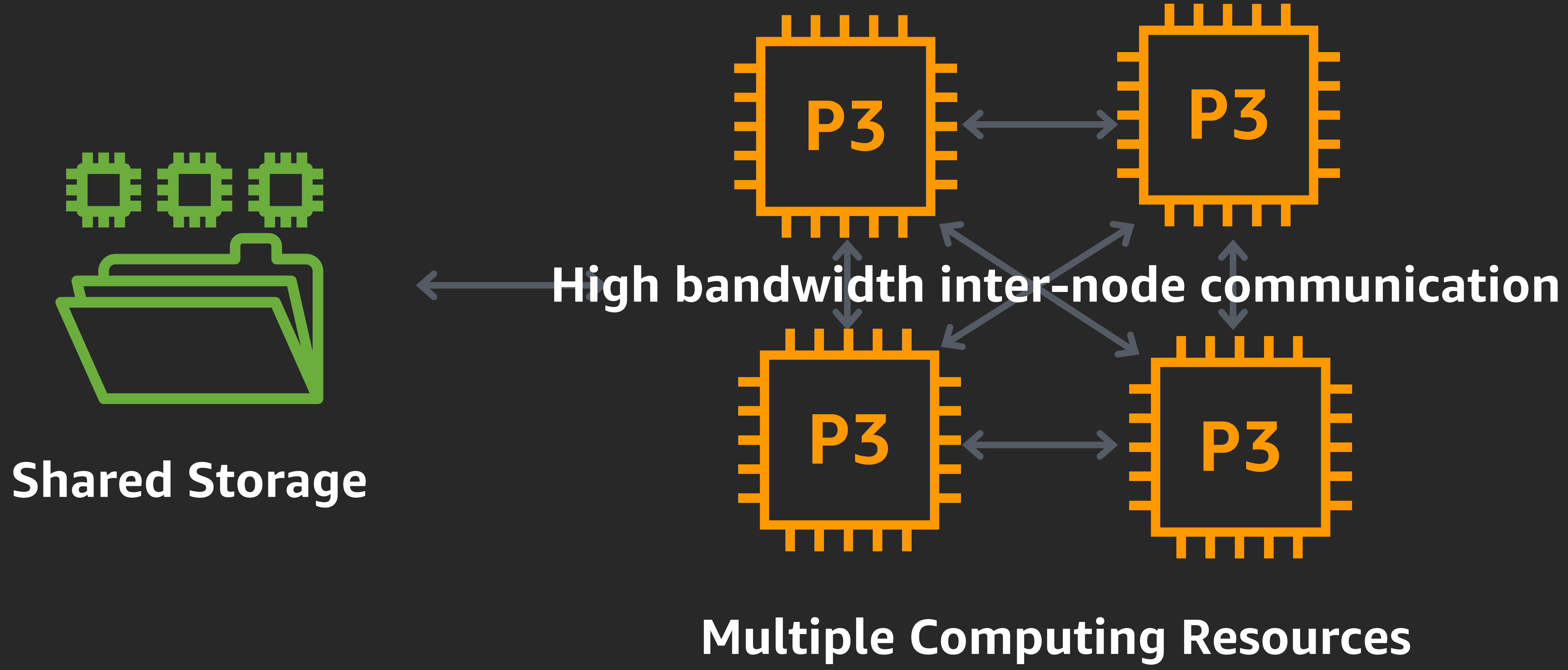
Options:

6.9% __더
6.1% __좋은
4.5% __더욱
2.2% __우리
2.0% __꼭
1.5% __모든
1.5% __정말
1.5% __건강
1.4% __작년보다
1.2% __모두
← Undo

SKT AIX에서 제공하고 있으며 UI는 lm-explorer를 기반으로 작성되었습니다. 문의 사항은 gogamza@sktair.com로 주세요.

Distributed training on AWS

Infra for distributed training - scale OUT



Tools for multi-GPU and distributed training

Deep learning framework

- TensorFlow - tf.distributed.Strategy
- PyTorch - DistributedDataParallel
- Apache MXNet - Parameter server

Toolkit for distributed training on top of deep learning frameworks

- All-reduce based distributed training framework, Horovod
- A deep learning optimization library, DeepSpeed

AWS ML services for distributed training

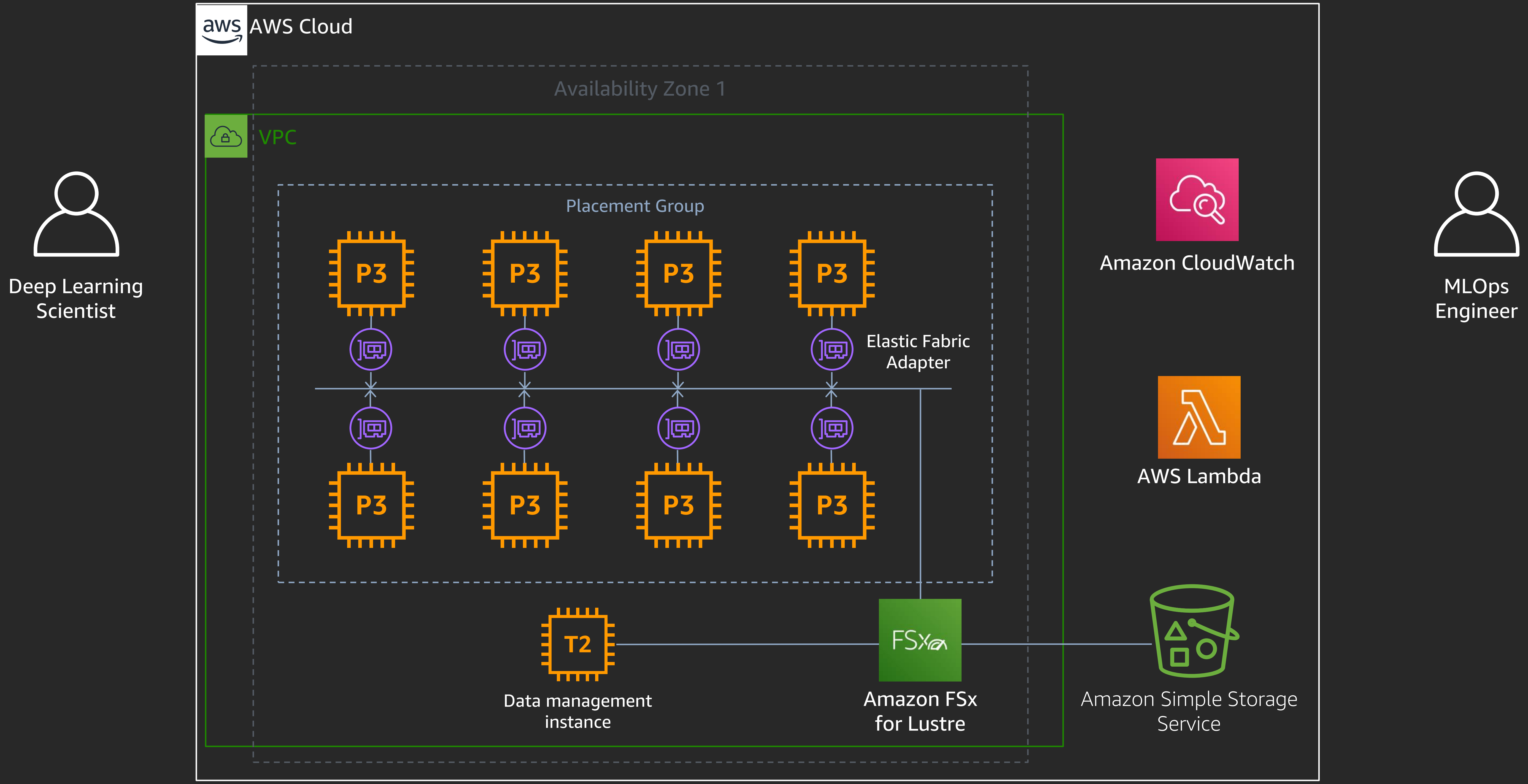
Amazon EC2

- Amazon EC2 P3dn.24xlarge for compute (NVIDIA V100 TensorCore GPU)
- Elastic Fabric Adapter for high speed network (100Gbps)
- Amazon FSx for Lustre for shared storage
- Deep Learning AMI
- EC2 Launch Templates or AWS Parallel Cluster for automation

Amazon SageMaker

- ml.p3dn.24xlarge ML instance for compute
- Elastic Fabric Adapter for high speed network
- Amazon S3 with SageMaker Pipe mode or Amazon FSx for Lustre

AWS architecture for KoGPT-2 training



What we, as a human team, did

CPU and GPU utilization monitoring

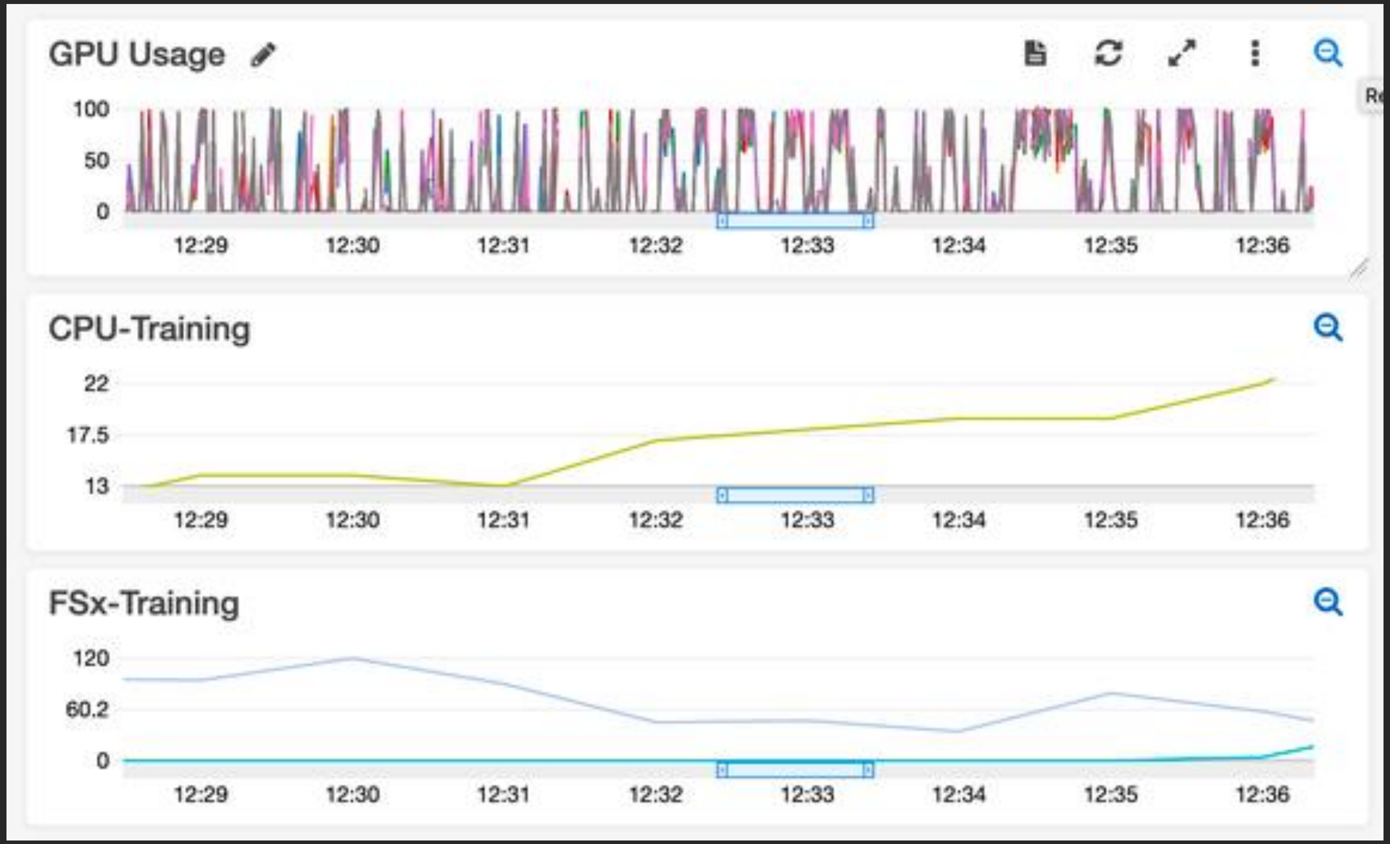
- GPU monitoring script
- Amazon CloudWatch

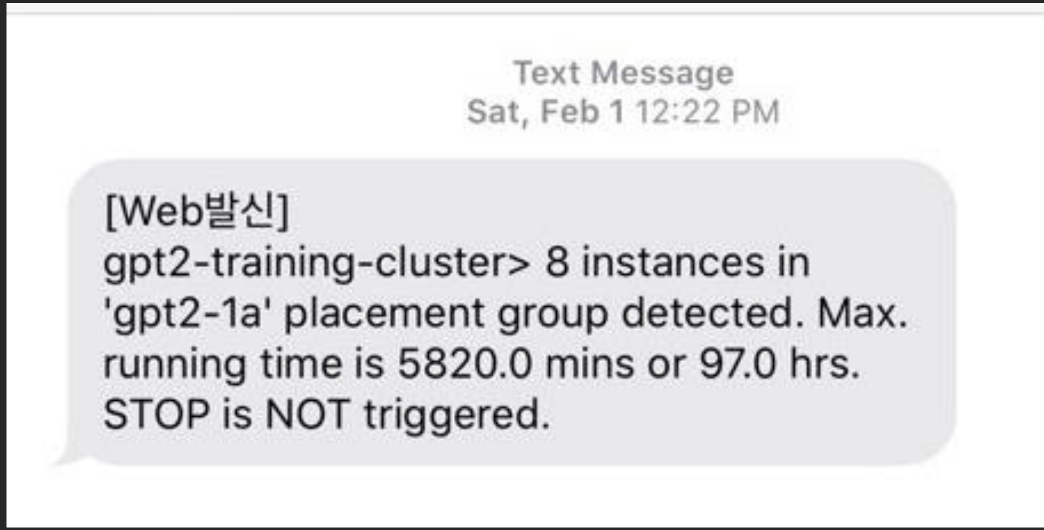
Bottleneck analysis

- Apache MXNet profiling
- Horovod timeline profiling

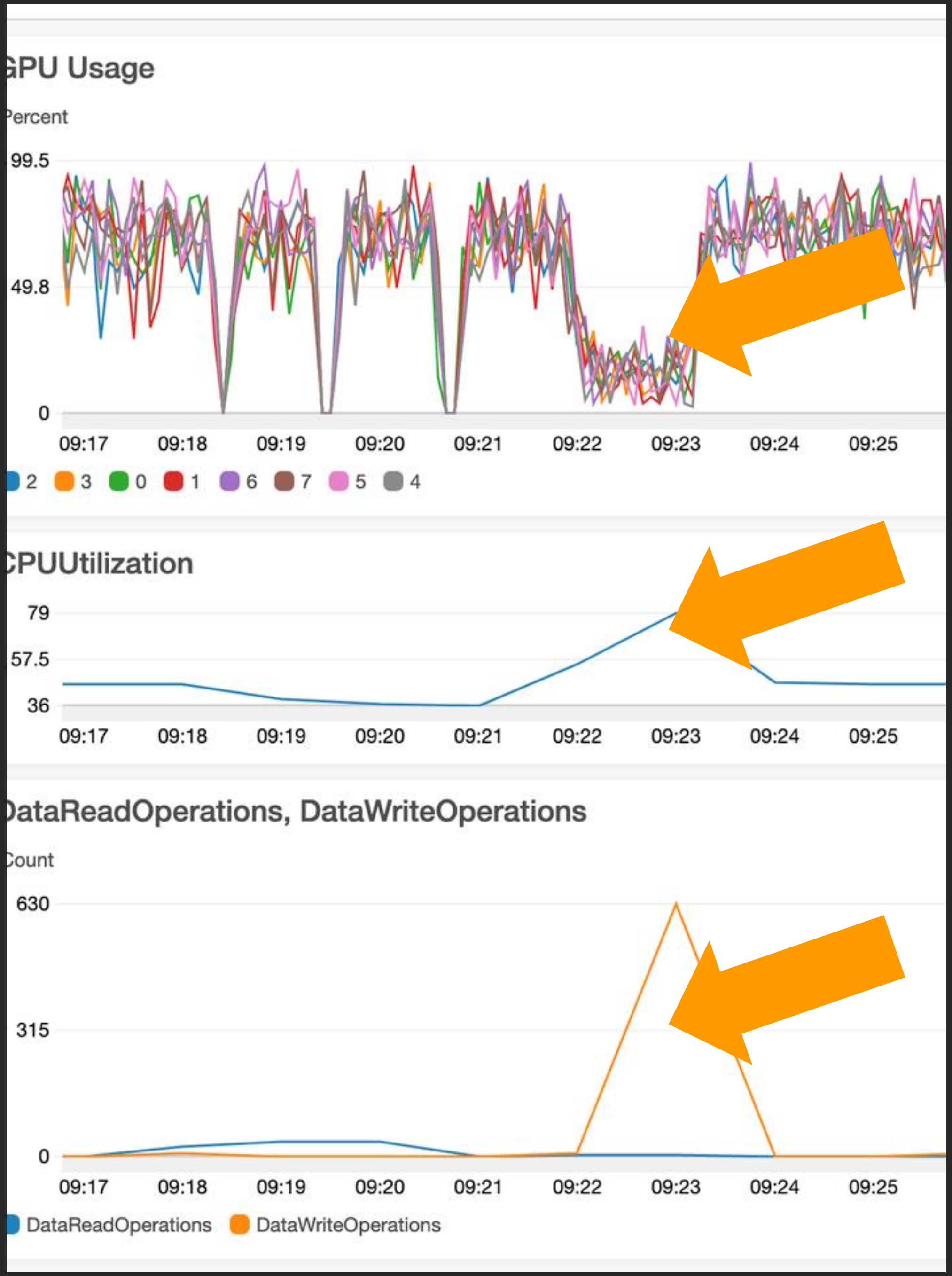
Code Review for better choices

Training running state check





Identifying bottleneck



GPU usage drop
+
CPU usage up
+
Disk write operation up

Removing
too frequent
checkpoint saving

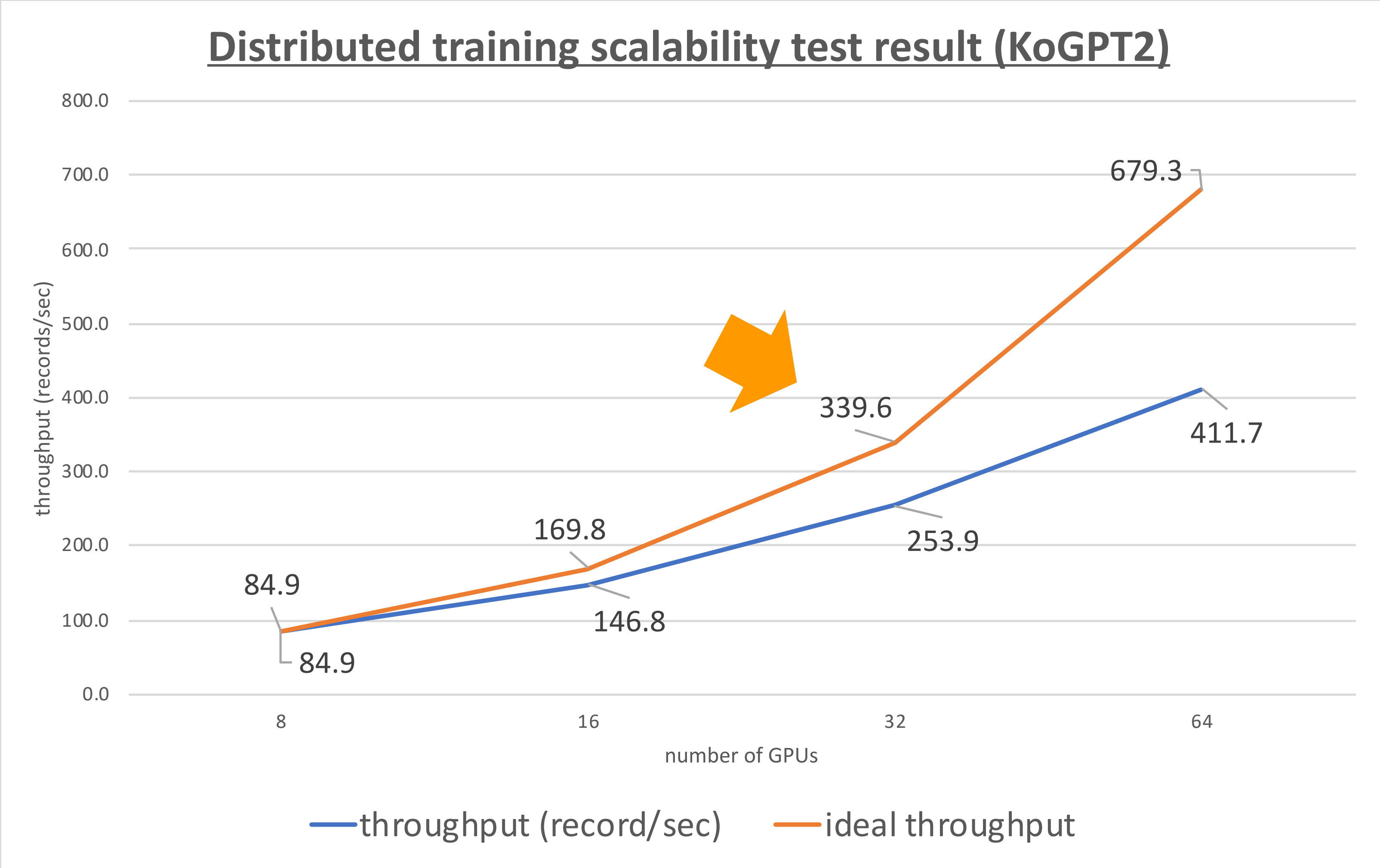
Tuning trials

- Activation function, GELU (+++)
- NVIDIA implementation of Adam, BERTAdam (+)
- Horovod option tuning
- Mixed precision training (+)
- *hvd.DistributedTrainer* instead of *hvd.DistributedOptimizer*
- Data transposing using GPU instead of using CPU (+)

Tuning result

# of node	throughput (rec/sec)	float32/16	throughput/ node	
1	53.4	float32	53.4	mxnet_cu101mkl-1.6.0b20191006
2	93.8	float32	46.9	
4	182.5	float32	45.6	
4	238.6	float16	59.7	
6	269.9	float32	45.0	
6	346.5	float16	57.8	
8	500.0	float16	62.5	
12	533.9	float16	44.5	
12	534.6	float16	44.6	tree
16	965.4	float16	60.3	tree
1	56.6	float32	56.6	mxnet_cu101mkl-1.6.0b20191230
1	75.4	float16	75.4	
1	81.0	float16	81.0	updated GELU

Final training decision



5 day training with 8 EC2 P3dn.24xlarge instances (64 Nvidia V100 GPUs)

Why not using Amazon SageMaker for training?

Yes!

Since all the training configuration and tuning have completed,
both training new pretrained models on new dataset
and
fine-tuning for downstream NLP tasks
on Amazon SageMaker is a desired option.

Inference on Amazon SageMaker

aws SUMMIT ONLINE

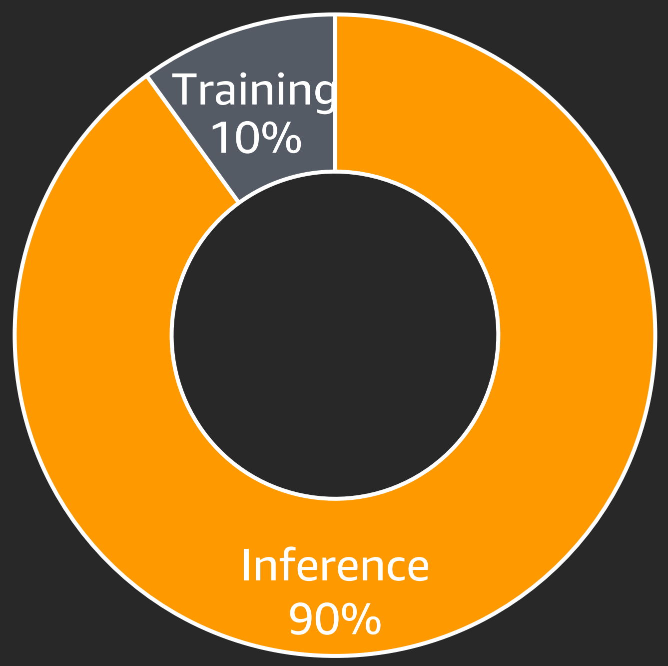
Amazon SageMaker:
Build, Train, and Deploy ML Models at Scale

Pre-built notebooks for common problems	Built-in, high performance algorithms	One-click training on the highest performing infrastructure	Model Optimization	One-click Deployment	Fully managed with auto-scaling for 75% less
Collect and prepare training data	Choose and optimize your ML algorithm	Set up and manage environments for training	Train and Tune ML Models	Deploy models in production	Scale and manage the production environment

Inference vs Training

Inference	Training
Usually run on a single input in real time	Requires high parallelism with large batch processing for higher throughput
Less compute/memory intensive	Compute/memory intensive
Integrated into the application stack workflows	Standalone, not integrated into an application stack
Runs on different devices at the edge and in the cloud	Run in the cloud
Runs all the time	Typically runs less frequently (train once, repeat infrequently)

The majority of the cost and complexity of Machine Learning (ML) in production is due to Inference

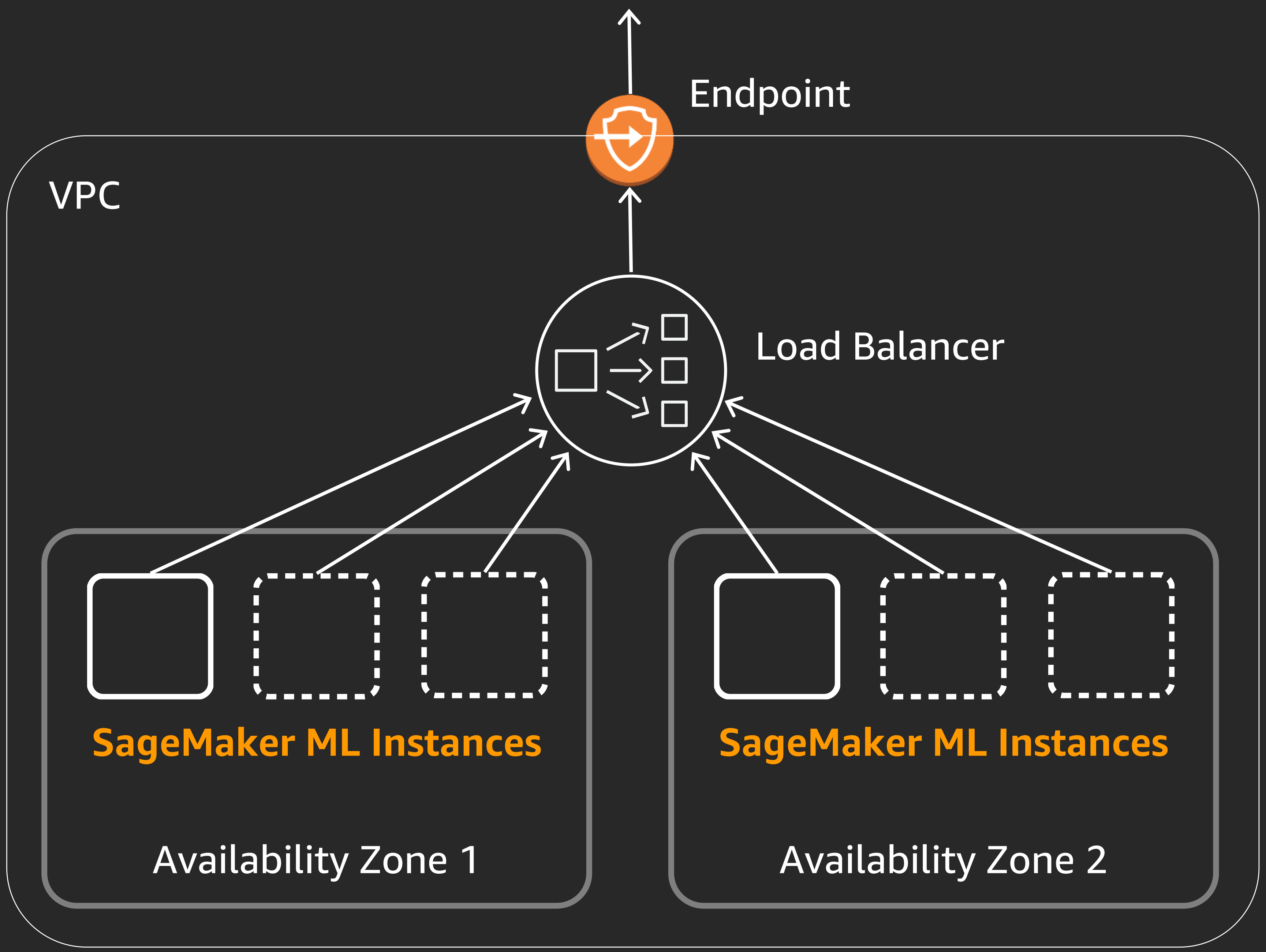


Amazon SageMaker – Instance Types for Inference

	Instances with CPUs				Instances with GPUs/Inference chips			
Instance Family	t family	m family	r family	c family	p family	g family	Inf1	Elastic Inference
Workload Type	Short jobs/ Notebooks	Standard CPU/ Memory ratio	Memory optimized	Compute optimized	Accelerated Computing- Training and Inference	Accelerated inference, smaller training jobs	Accelerated Computing - Inference	Cost-effective Inference Accelerators
	t3.2xlarge 8 vCPU 32 Mem	m5.2xlarge 8 vCPU 32 Mem	r5.2xlarge 8 vCPU 64 Mem	c5.2xlarge 8 vCPU 16 Mem	p3.2xlarge 8 vCPU 61 Mem 1xV100 GPU	g4dn.2xlarge 8 vCPU 32 Mem 1xT4 GPU		

https://aws.amazon.com/sagemaker/pricing/instance-types/

SageMaker Endpoint - Scalable and Highly Available



- Deploy more than one instance behind an inference endpoint for high availability
- Enable automatic scaling with a predefined metric or a custom metric
- You can manually change the instance number and type without incurring downtime
- Set Amazon CloudWatch alarms on availability and error rates

Deploying in SageMaker at Any Scale

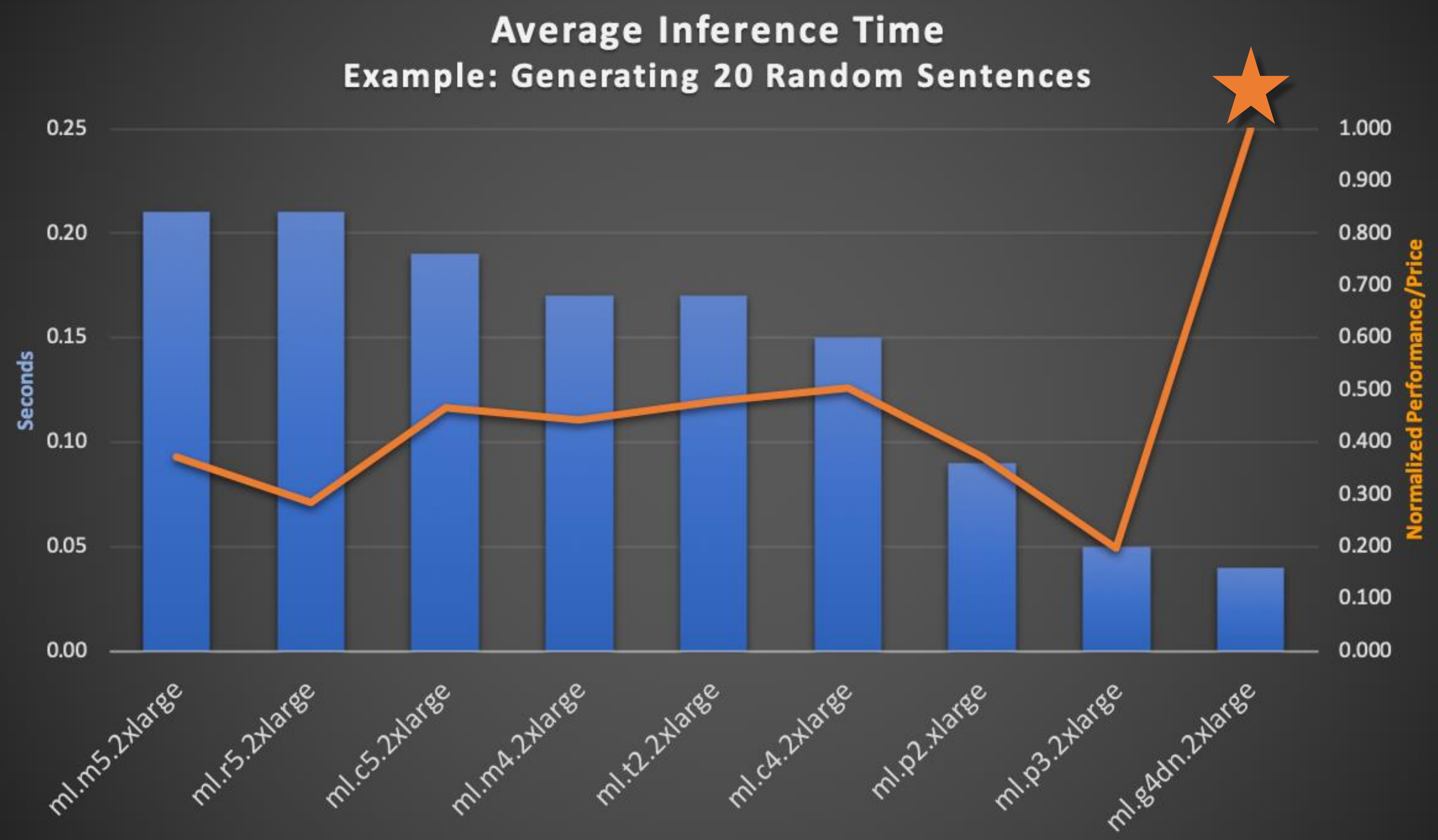
```
sagemaker_session = sagemaker.Session()
role = get_execution_role()

model_data = 's3://<your bucket name>/gpt2-model/model.tar.gz'
entry_point = './gpt2-inference.py'

mxnet_model = MXNetModel(model_data=model_data,
              role=role,
              entry_point=entry_point,
              py_version='py3',
              framework_version='1.6.0',
              image='<AWS account id>.dkr.ecr.<AWS region>.amazonaws.com/kogpt2:latest'
)

# HTTPS endpoint backed by 16 instances, multi-AZ and load-balanced
predictor = mxnet_model.deploy(instance_type='ml.c5.xlarge', initial_instance_count=16)
```

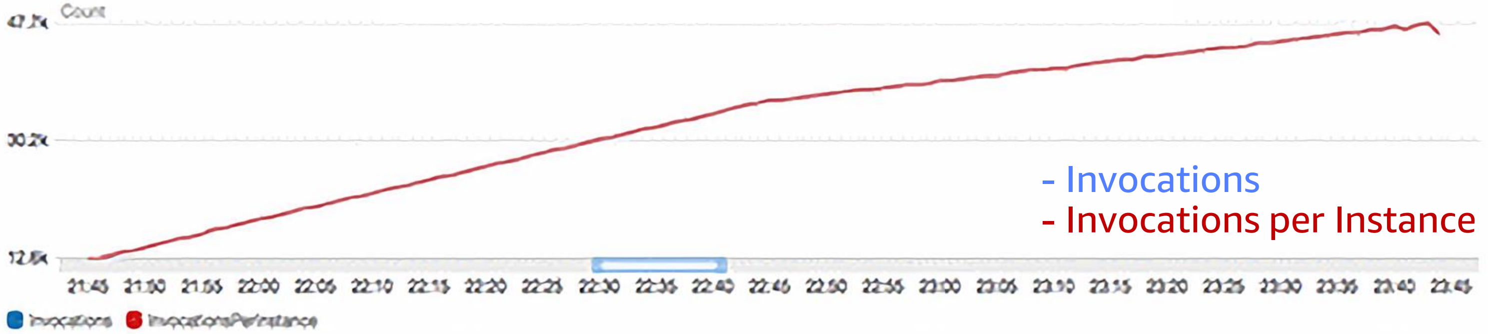
```python
data_list = [
          '2019년 한해를 보내며,',
          '오늘 5월 13일 날씨는 아침에 전국 대부분의 지역이 맑은 날씨를 보이는 가운데',
          '앞서 간밤 미국 뉴욕증시에서는 다우존스 30 산업평균지수가',
          '류현진 선수는 올 시즌을',
          '프리미어리그 사무국은 28일 공식 홈페이지를 통해',
          '아마존닷컴은 신종 코로나 바이러스에 대비하려는 고객들을 위해 마스크, 소독제, 물티슈 등 생필품 가격을
          '도널드 트럼프 미국 대통령은 현지시간으로 25일 문재인 대통령과의 전날 전화 통화에서',
          '앞서 조 회장이 작년 크리스마스 당일 어머니',
          '한편 미국에서는 코로나바이러스에 감염된 환자들의',
          '유치원과 초·중·고등학교 교사 73%가 개학을 4월 6일 이후로',
          '제 뒤로 보시는 것처럼 이곳은 연분홍색 벚꽃이 활짝 피었고',
          '기업들이 빠르게 재택근무 환경을 구축하면서 아마존웹서비스의 주가',
          '플라시도 도밍고는 세계적 테너로 명성을 얻었지만, 지난해',
          '2000년대 초반 전성기를 맞았던 그는 최근 몇년간',
          '방송에서 새로운 셰프로 등장한 그는 "아이들 건강을 위해 집에서',
          '오는 31일부터 한부모 노동자가 육아휴직을 사용할 경우',
          '이는 코로나19 사태로 센터가 임시휴관에 들어감에 따라',
          '재즈와 블루스 연주에서 독보적인 위치에 있는 정씨를 비롯해',
          '록그룹 퀸의 기타리스트 브라이언 메이가 호주 산불피해로 치료중인 코알라를 위해',
          '2020년 프로야구 개막이 4월 중으로 연기된 가운데 두산'
]
```


Average Inference Time
Example: Generating 20 Random Sentences
Seconds
0.25
0.20
0.15
0.10
0.05
0.00
Normalized Performance/Price
1.000
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
ml.m5.2xlarge
ml.r5.2xlarge
ml.c5.2xlarge
ml.m4.2xlarge
ml.t2.2xlarge
ml.c4.2xlarge
ml.p2.xlarge
ml.p3.2xlarge
ml.g4dn.2xlarge

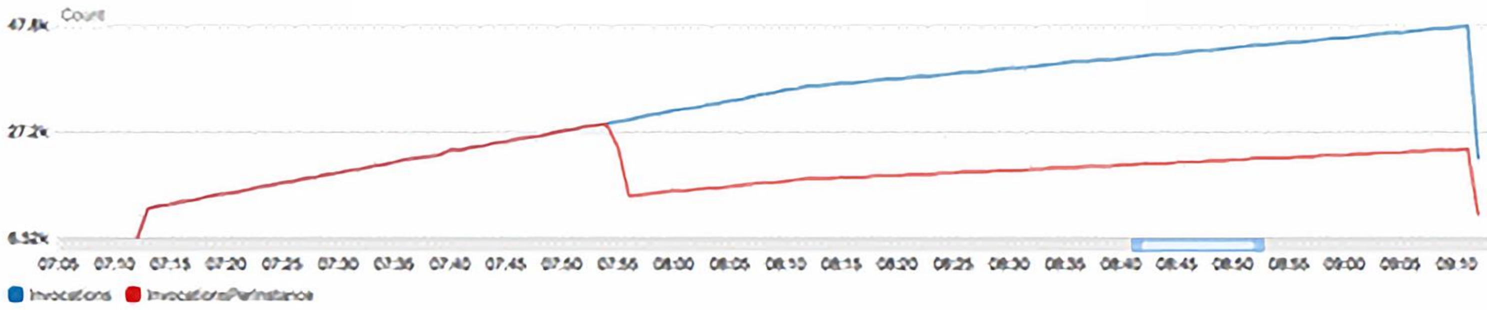
Auto Scaling Your Endpoint

Scaling policy : Target value of 25000 (per min) for `SageMakerVariantInvocationsPerInstance`

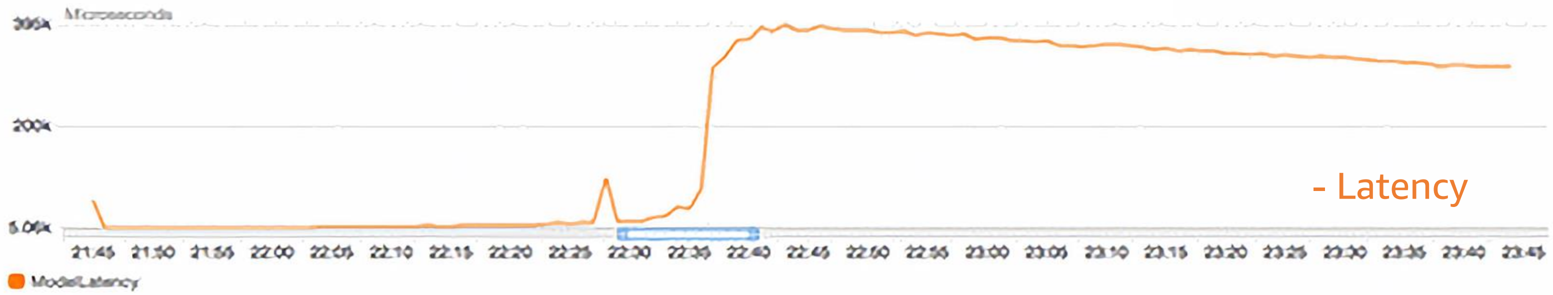
Auto Scaling OFF (1 instance)



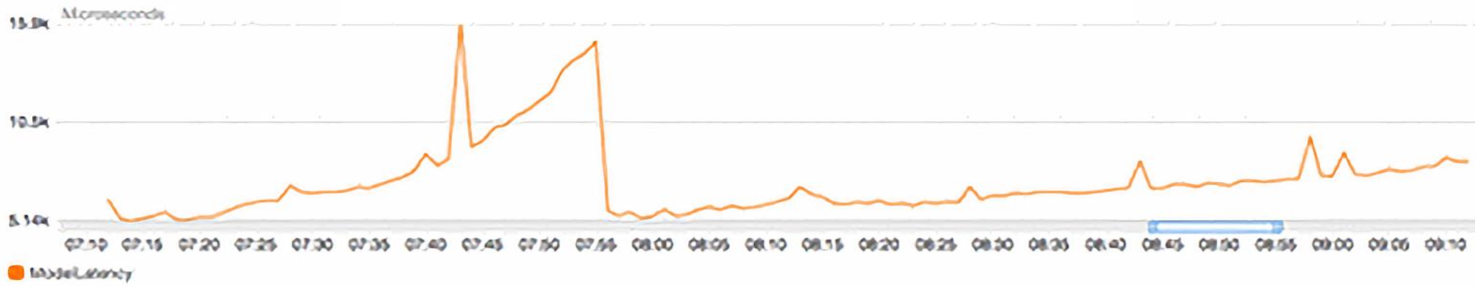
Auto Scaling ON



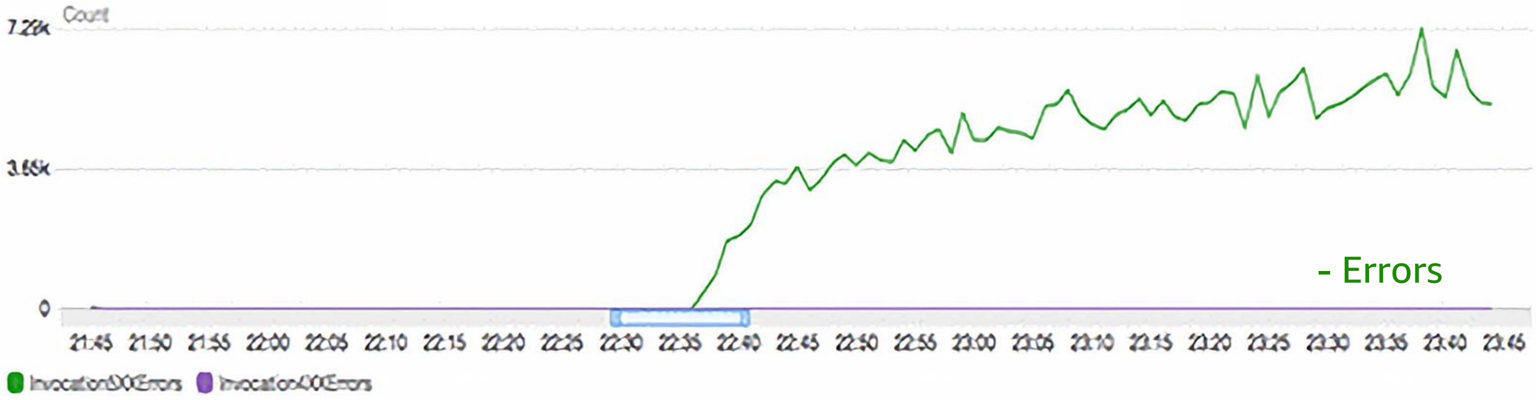
Auto Scaling OFF (1 instance)



Auto Scaling ON



Auto Scaling OFF (1 instance)



Auto Scaling ON

References

KoGPT-2 : https://github.com/SKT-AI/KoGPT2

AWS Samples GitHub : https://github.com/aws-samples

AWS Korea Blog : https://aws.amazon.com/ko/blogs/korea/

Machine Learning on AWS : https://ml.aws

Amazon SageMaker : https://aws.amazon.com/sagemaker/

GluonNLP : https://gluon-nlp.mxnet.io/

Amazon ML Solutions Lab : https://aws.amazon.com/ml-solutions-lab/

여러분의 소중한 피드백을 기다립니다!

강연 평가 및 설문 조사에 참여해 주세요.

감사합니다

AWS SUMMIT ONLINE