거대 인공 지능 모델을 위한 고효율 컴퓨팅 인프라

Large AI Models and Computing Challenges

김홍연 (kimhy@etri.re.kr)
한국전자통신연구원

CONTENTS

01 거대 AI 모델을 위한 컴퓨팅

02 고효율 AI 컴퓨팅 기술 연구

03 추진 체계 및 공개 SW 전략

01

거대 AI 모델을 위한 컴퓨팅

01 거대 AI 모델 발전

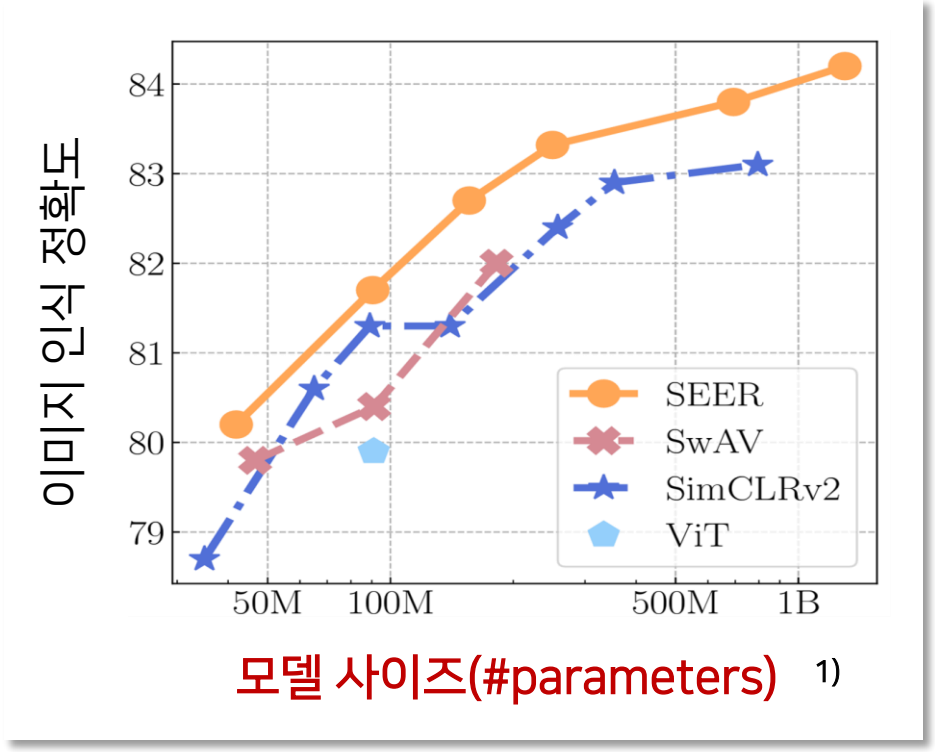
- **거대 AI 모델 출현 및 발전 가속화**
  - GPT-3(OpenAI)
    - 트랜스포머 모델, 1,750억개 파라미터, 퓨/제로샷(few/zero-shot) 전이 학습을 통해 하나의 모델로 다양한 태스크 수행 가능 (출처: OpenAI)
  - PaLM(Google)
    - 트랜스포머 모델, 5,400억 파라미터, 6종 학습 데이터(소셜 미디어, 웹페이지, 책, Github, Wikipedia, News), 프롬프트 튜닝을 사용하여 단일 모델로 프로그래밍, 수학 문제 해결, 다국어 번역, 논리적 추론 가능 (출처: Google)
  - Gato(DeepMind)
    - 멀티 모달 멀티 태스크 트랜스포머 모델로 아타리 게임, 이미지 캡션, 채팅, 블록 쌓기 등 600여 가지 태스크 수행 가능한 범용 에이전트 출시 (출처: DeepMind)

→ **보상, 모델 거대화, 학습 데이터 다양화**로 AGI 달성 가능 주장

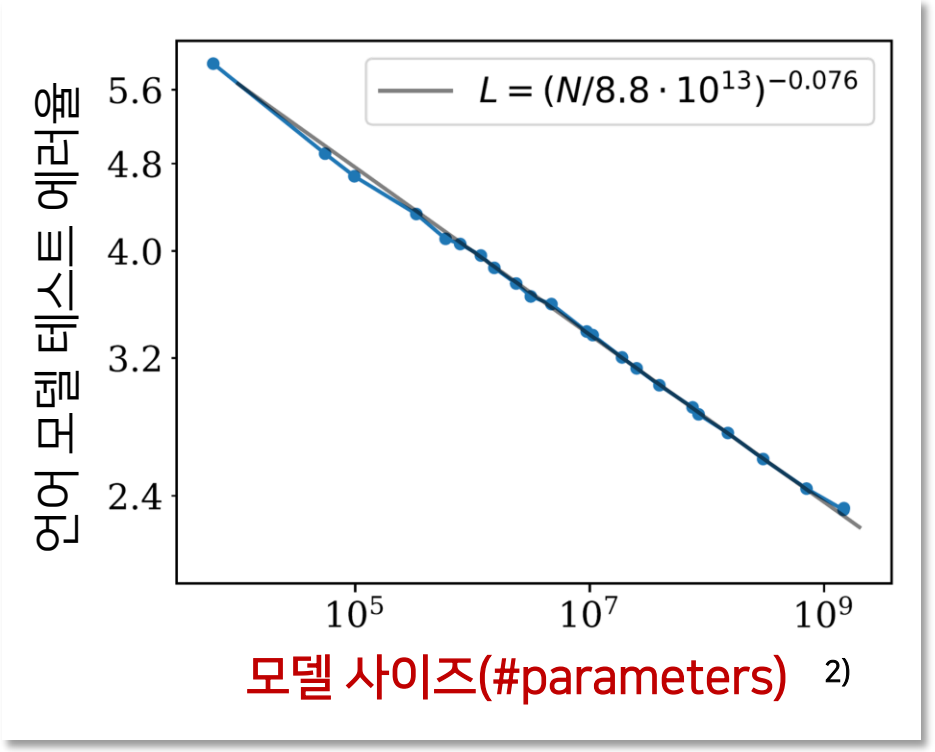
(출처: Reward is Enough, DeepMind, AI Journal '21)

01 AI 모델 거대화 이유

- Scaling Law: 모델 사이즈가 클수록 인공지능의 정확도 향상



1)

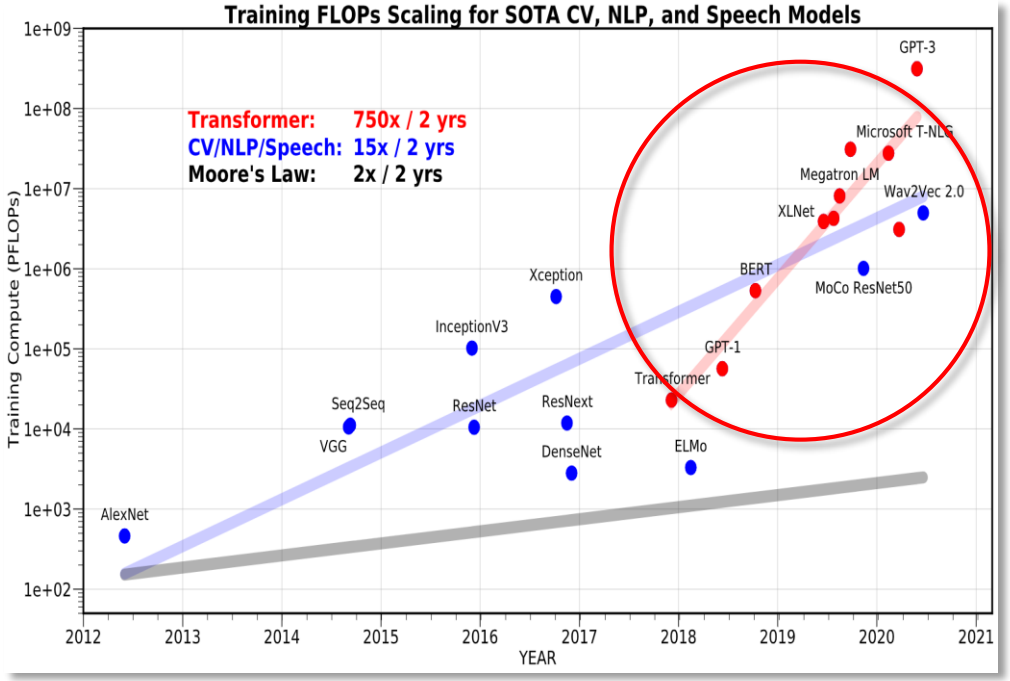


2)

1) Goyal et al., *Self-supervised Pretraining of Visual Features in the Wild,* arXiv 2021.
2) Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv 2020.

01 초거대 AI 모델 발전의 한계

- AI 모델 크기 증가 속도 → 무어의 법칙(Moore's Law) 상회 → **성능 장벽**
- 초거대 AI 모델 학습 → 대규모 에너지 필요 → **에너지 장벽**



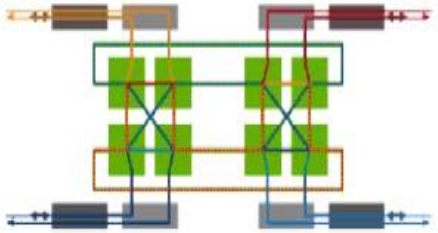
성능 장벽과 에너지 장벽을 해결해야 초거대 AI 모델 발전 가능

01 초거대 AI 모델 발전의 한계

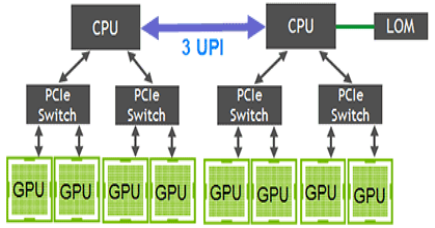
- GPU 클러스터/클라우드 비용(TCO) 부담 폭증 → **비용 장벽**

	Megatron-LM 사례 (DGX A100 하이퍼 클러스터 활용)			
목표 효율	파라메터수	1조	1,750억	17억
	유효 성능 (Pflops)	502	143.80	4
	컴퓨팅 효율	52%	45%	44%
가속기 성능스펙	GPU	A100	A100	A100
	TF16 성능 (TFlops)	312	312	312
요구성능	구축 성능 (PFlops)	965.4	319.6	10
	총 GPU 수	3,072	1,024	32
	총 서버 수 (8 GPUs/server)	384	128	4
비용	개별 GPU 가격 (만원)	1.5	1.5	1500
	개별 서버 가격 (억원)	4.6	4.6	4.6
	추정 가격 (억원)	1,766	589	18

DGX A100 기반 하이퍼 클러스터 비용 비교
(출처: Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM, SC21, 2021)



p3.16xlarge GPU Instance
(V100 8GPU, NVLink)



G5.48xlarge GPU Instance
(A10G 8GPU, PCIe)

GPU 인스턴스	GPU 개수	GPU 연결망	시간당 비용	6개월 비용 (개발, 시행착오, 학습 포함)
g5.48xlarge	32	PCI-E	$16.29	6.96 억원 ($562,982)
p3.16xlarge	32	NVLink	$24.48 ($7.34 spot)	10.56 억원 ($846,029)

17억 파라메터 규모를 위한 Amazon EC2 인스턴스 종류 및 비용 비교
(출처: https://aws.amazon.com/ko/ec2/instance-types)

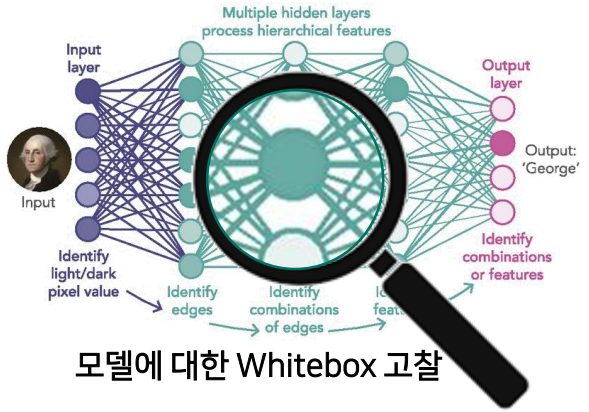
막대한 컴퓨팅의 종속성 및 비용 부담 또한 초거대 AI 모델 발전의 걸림돌

01 한계 돌파 기술: 시스템적 최적화 vs. 비시스템적 최적화

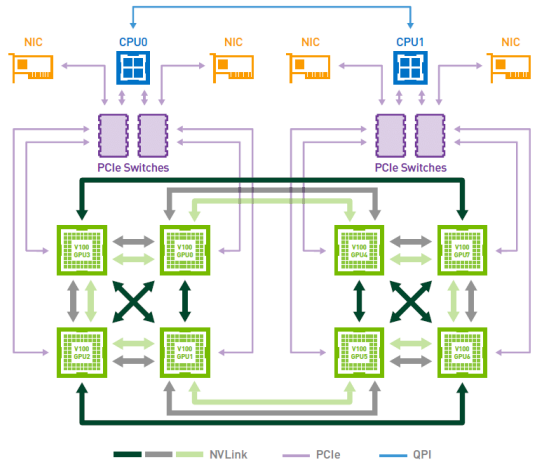
- 기술 진화 사이클 = 기능(정확도) ↑ → 절대 성능 ↑ → 효율(범용성) ↑
  - 초기 Model Parallelism 효율 5% → SOTA 50% 수준
  - 프로세서/서버간 통신 오버헤드가 제약조건

- 비시스템적 (모델 구조, 학습, 추론 방법 변경) 최적화 기법
  - Switch Transformer (2021, Google)
  - ALBERT (ICLR 2020)
  - Progressively Stacking (PMLR 2019)
  - Progressive Layer Dropping (NeruIPS 2020)
  - Faster Transformer (GTC 2020, NVIDIA)

- 시스템적 최적화 기법
  - Varuna (Eurosys 2022, Microsoft)
  - DeepSpeed (ATC 2021, Microsoft)
  - Megatron-LM (GTC 2020, NVIDIA)



모델에 대한 Whitebox 고찰



컴퓨팅 특성 활용 최적화
(예: GPU-awareness)

02

고효율 AI 컴퓨팅 기술 연구

02 문제 정의

모델

NVIDIA
Megatron-LM 전용
Bert, T5, GPT

HuggingFace
83,000 Models

EleutherAI
GPT-NeoX
GPT-J
GPT-Neo

문제1) 모델, 프레임워크 종속성

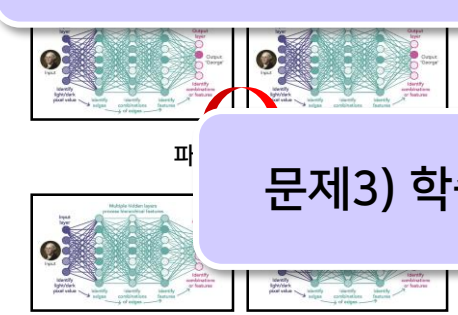
프레임워크 및 병렬화 기술

deepspeed

PyTorch GPipe

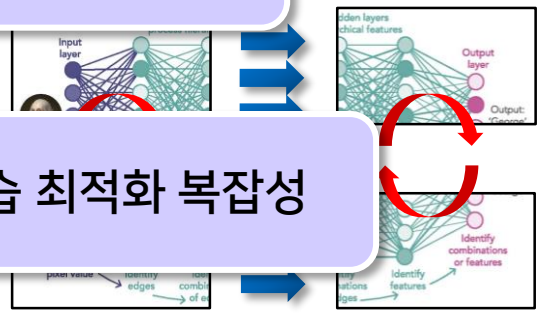
TensorFlow

NVIDIA
Megatron-LM

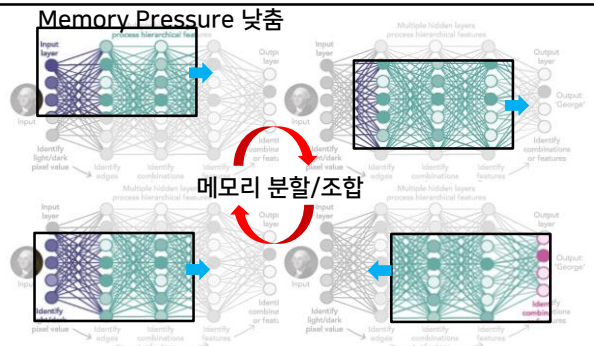


데이터 병렬화
(모델 크기가 GPU 용량 초과시는 불가능)

문제3) 학습 최적화 복잡성



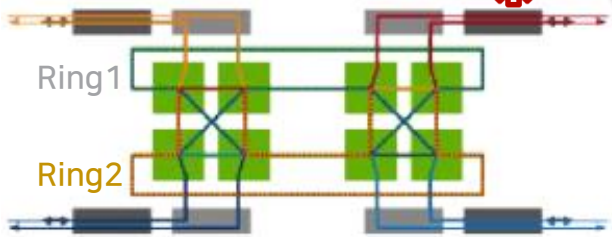
모델 (텐서, 파이프라인) 병렬화
(MS Pipedream, Megatron-LM 등)



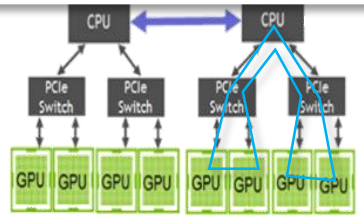
메모리 최적화
(NVIDIA vDNN, Microsoft Zero 등)

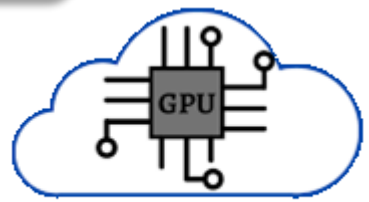
문제2) HW 및 Architecture 종속성

HW



하이퍼 클러스터의 HW Defined 고성능 통신으로 인한 높은 효율

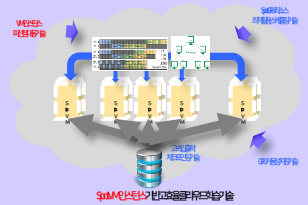




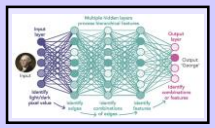
범용 클러스터 또는 클라우드 GPU의 저대역 통신으로 인한 낮은 컴퓨팅 효율

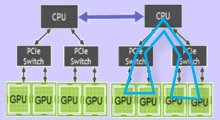
02 연구 개발 범위

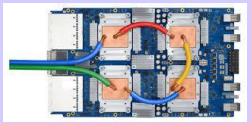
클라우드 효율 향상 기술

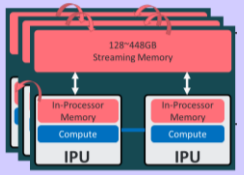


종속성 완화 기술









ML Applications
Computer Vision | Speech | Translation | Autonomous Driving

ML Model
ResNet | GPT | BERT | Wu Dao

AI Cloud
Google Vertex AI | MS AzureML | Amazon SageMaker | KubeFlow

Distributed ML Framework
Megatron-LM | DeepSpeed | SageMaker | Varuna | ALPA

ML Framework
TensorFlow | PyTorch | JAX | Caffe | MXNet

Compiler
ALPA | XLA | Glow | TVM

Auto Parallelization
PyTorch-Lightning | GSPMD | FlexFlow

Optimization Library
CUDA | OpenCL | cuBLAS | OpenBLAS

Collective Comm. Library
NCCL | MPI

Hypervisor
KVM | VMware

OS
Docker | Linux | Windows

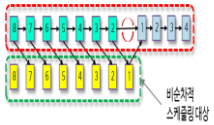
Monitoring/Management
OpenStack | Kubeflow | Kubernetes

Accelerator
GPU | TPU | IPU | NPU | FPGA

Interconnect
NVLink | IB | PCIe | CXL

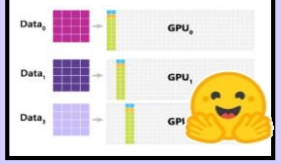
학습 효율 향상 기술

고효율 학습 병렬화



성능효율적 커널 퓨전 기능 제공

메모리 효율화

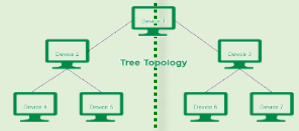


동적 그래프 기반 파이프라인 병렬화

02 연구 개발 로드맵

1단계 (22년~23년)
학습 효율 향상 기술 개발

고효율 병렬 학습 기술



저대역 네트워크 친화적 학습 통신 기술

고효율 병렬학습 프레임워크
(v1.0, Bumblebee)

2단계 (24년)
종속성 완화·사업화 기술 개발

프레임워크 생태계
확장 및 국산화

PyTorch
TensorFlow
Tunib OSLO

Spot 인스탄스
클라우드 학습

이종 가속기 지원

IPU

TPU

기술 고도화
(I/O, 메모리)

고효율 병렬학습
프레임워크
(v2.0, Jazz)

3단계 (25년)
기술 검증 및 사업화

기술 검증 및
사업화 기술 지원

ETRI

국내외
레퍼런스 확보 및
사업화

TUNiB

GRAPHCORE

최우수 학회
논문 발표

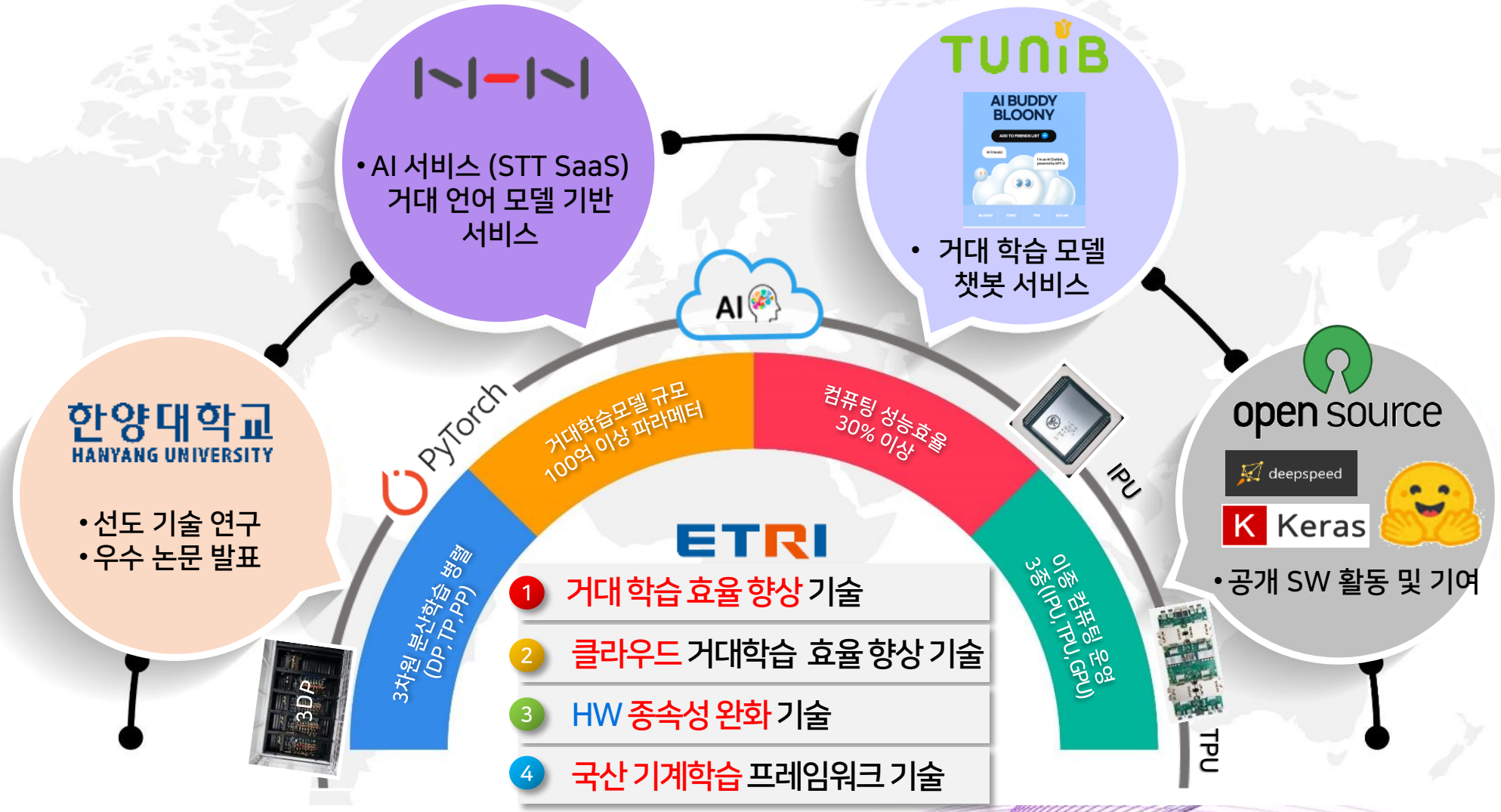
한양대학교

고효율 병렬학습 프레임워크
(v3.0, Optimus Prime)

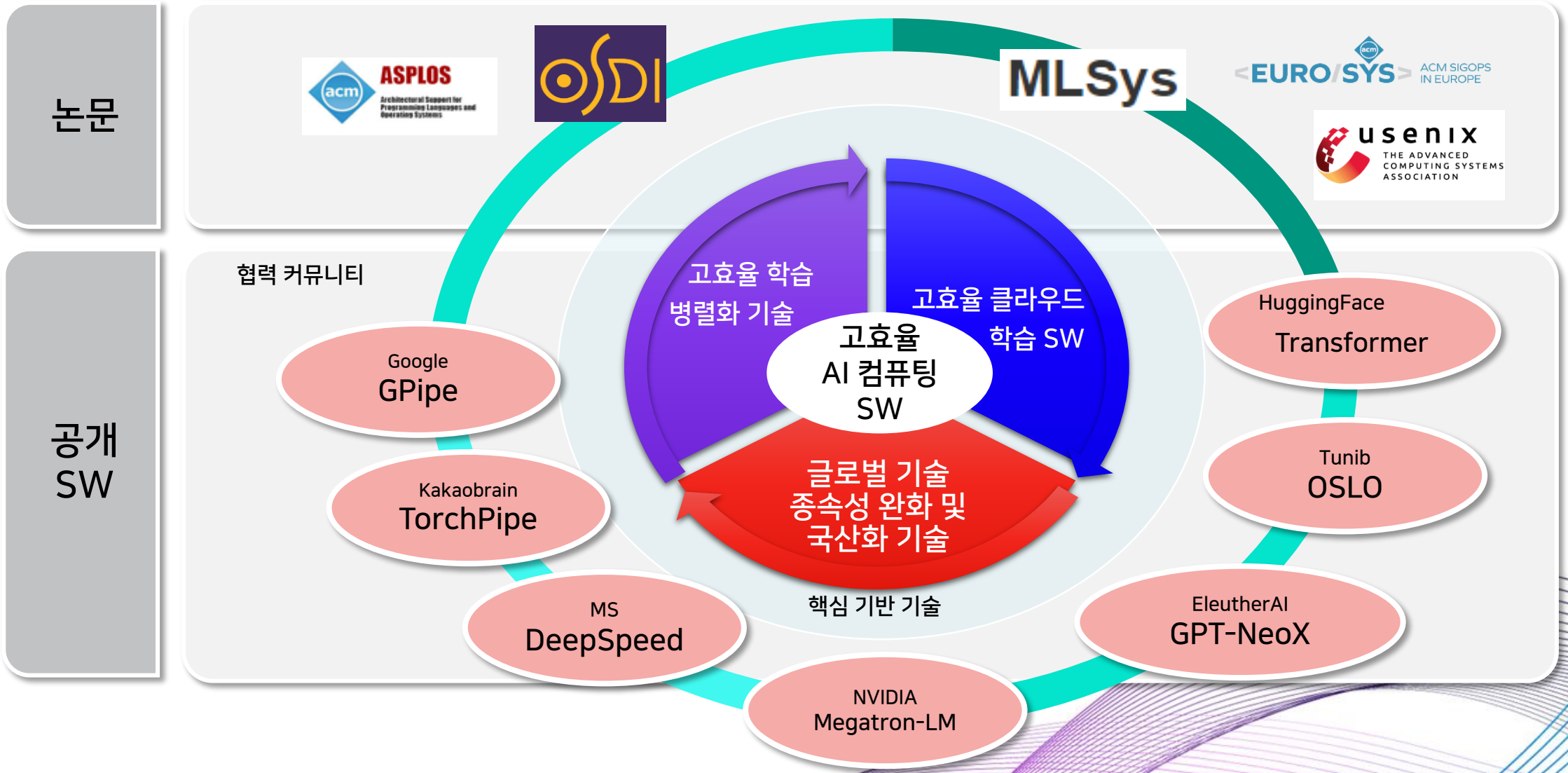
03

추진 체계 및 공개 SW 전략

03 추진 체계


• AI 서비스 (STT SaaS)
거대 언어 모델 기반
서비스
TUNiB
AI BUDDY
BLOONY
• 거대 학습 모델
챗봇 서비스
한양대학교
HANYANG UNIVERSITY
• 선도 기술 연구
• 우수 논문 발표
AI
PyTorch
거대학습모델 규모
100억 이상 파라메터
컴퓨팅 성능효율
30% 이상
3차원 분산학습 병렬
(DP, TP, PP)
이종 컴퓨팅 지원
3종(IPU, TPU, GPU)
IPU
TPU
3DP
ETRI
1 거대 학습 효율 향상 기술
2 클라우드 거대학습 효율 향상 기술
3 HW 종속성 완화 기술
4 국산 기계학습 프레임워크 기술
open source
deepspeed
Keras
• 공개 SW 활동 및 기여

03 공개SW 전략


논문
공개 SW
ASPLOS
OSDI
MLSys
EUROSYS ACM SIGOPS IN EUROPE
usenix THE ADVANCED COMPUTING SYSTEMS ASSOCIATION
협력 커뮤니티
고효율 학습 병렬화 기술
고효율 클라우드 학습 SW
고효율 AI 컴퓨팅 SW
글로벌 기술 종속성 완화 및 국산화 기술
핵심 기반 기술
Google GPipe
Kakaobrain TorchPipe
MS DeepSpeed
NVIDIA Megatron-LM
EleutherAI GPT-NeoX
Tunib OSLO
HuggingFace Transformer

감사합니다

거대 인공 지능 모델을 위한 고효율 컴퓨팅 인프라