

# **Design of Efficient Virtual Desktop Infrastructure based on Super Resolution And NPU**

---

\*Joo Hyoung Cha, Young Woon Woo

Dong-eui University



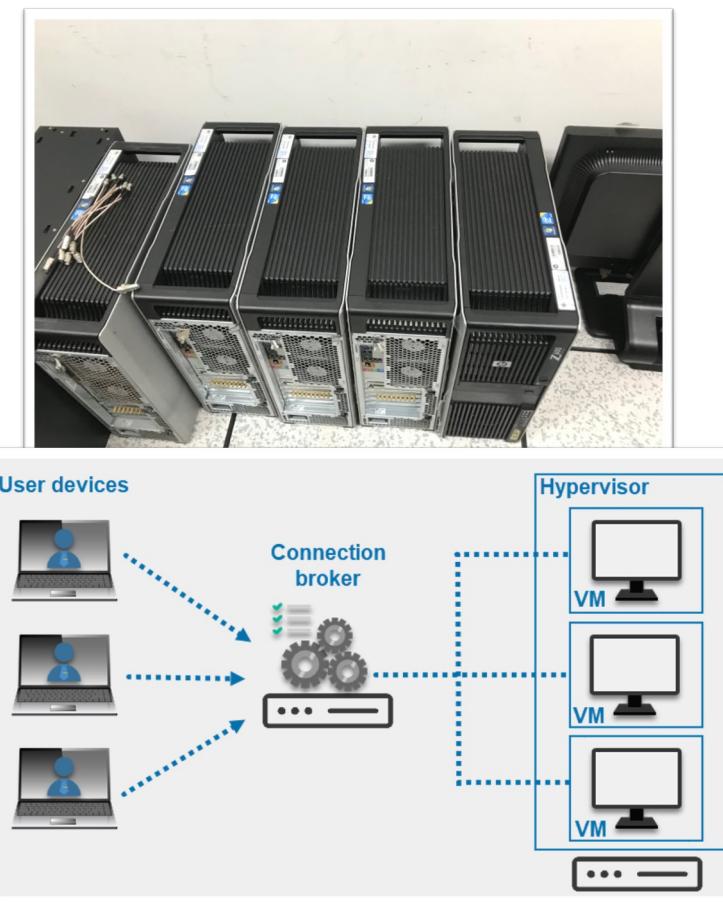
# CONTENTS

- 01 연구 배경
- 02 기존 연구 사례
- 03 시스템 설계
- 04 시스템 구성
- 05 성능 분석
- 06 결론

2023년 인공지능 및 응용 워크숍

# 01 연구 목적

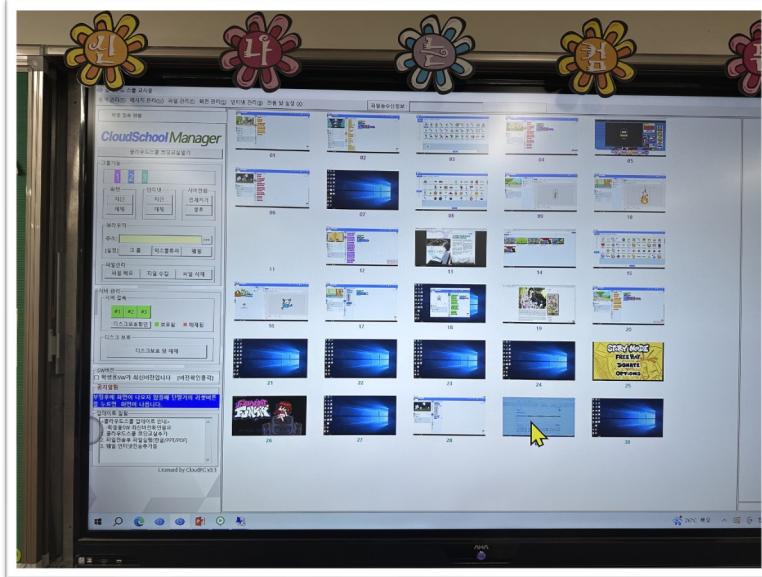
## 사용자 보안과 통신 비용이 고려된 DaaS(Desktop As A Service) 시스템



1. 공공기관 및 기업에서 컴퓨터 구매 후 최소 8년 이상 운용하여야 함.
2. 노후화가 쉽게 발생하고, 시간이 지남에 따라 작업 속도가 느려짐.
3. 따라서 성능이 높은 1대의 시스템을 구매 후, 장기간 사용하는 것이 중요함.
4. 그러나, 높은 비용으로 인하여 접근성이 낮음.
5. 이에, 유연한 성능, 확장성이 높은 Private Cloud를 도입이 필요함.
6. 기존 연구에는 통신 비용과 엔드 유저의 보안이 고려가 되지 않음.
7. 본 연구는 통신 비용과 사용자 인증 시스템을 통합된 Virtual Desktop Infrastructure를 제안함.

# 01 기존 사례

## 과정초등학교에 적용된 VM 기반의 DaaS(Desktop As A Service) 환경



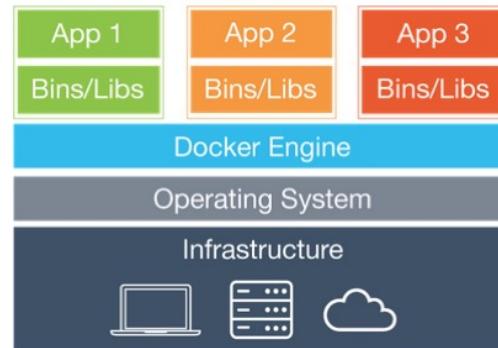
- Pros. 컴퓨터 단가를 절감 및 통제된 환경 제공
- Cons. 서버의 성능(20코어, 64GB)은 높으나 제공되는 성능이 낮음 (1코어, 2GB)

## 02 제안하는 인프라 제어를 위한 솔루션

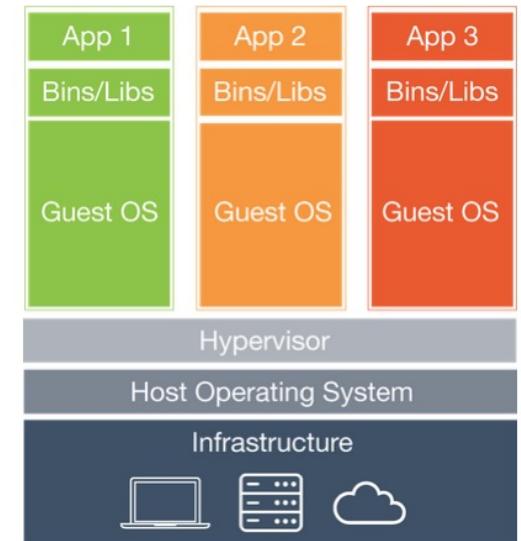
### Kubernetes(Container)와 OpenStack(VM)

1. OpenStack(VM) : 하드웨어 레벨의 자원 할당을 통한 가상화
2. Kubernetes(Container) : Kernel을 공유하고 OS를 가상화함.
3. 사용자가 할당된 자원 이상을 이용하기 위해 VM 보다 Container가 선호됨.
4. 따라서, 외부 서비스로 운용한다면 OpenStack 기반이지만, 사내에서 효율적인 시스템은 Container 기반이 효율적임.
5. 따라서, Container 기반의 VDI 인프라 제어를 제안함.

### Kubernetes



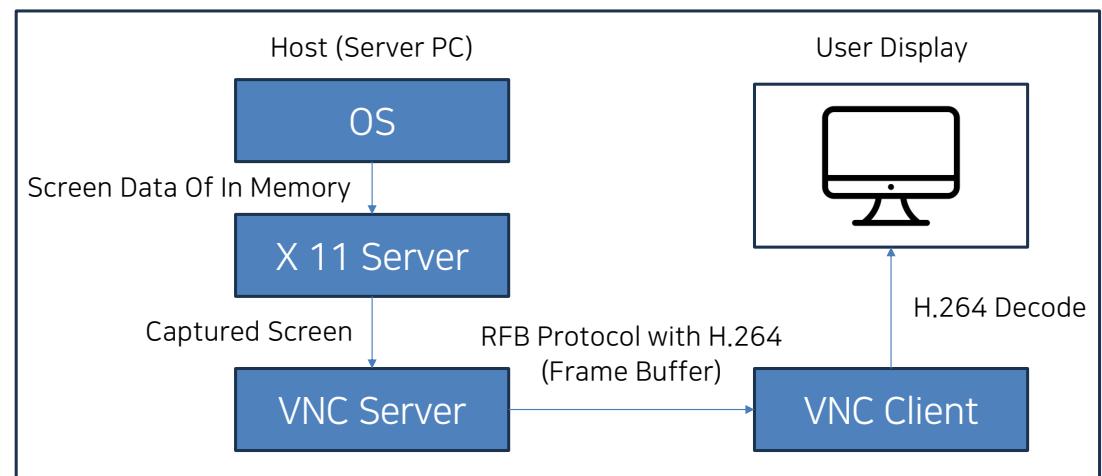
### OpenStack



# 01 원격 제어

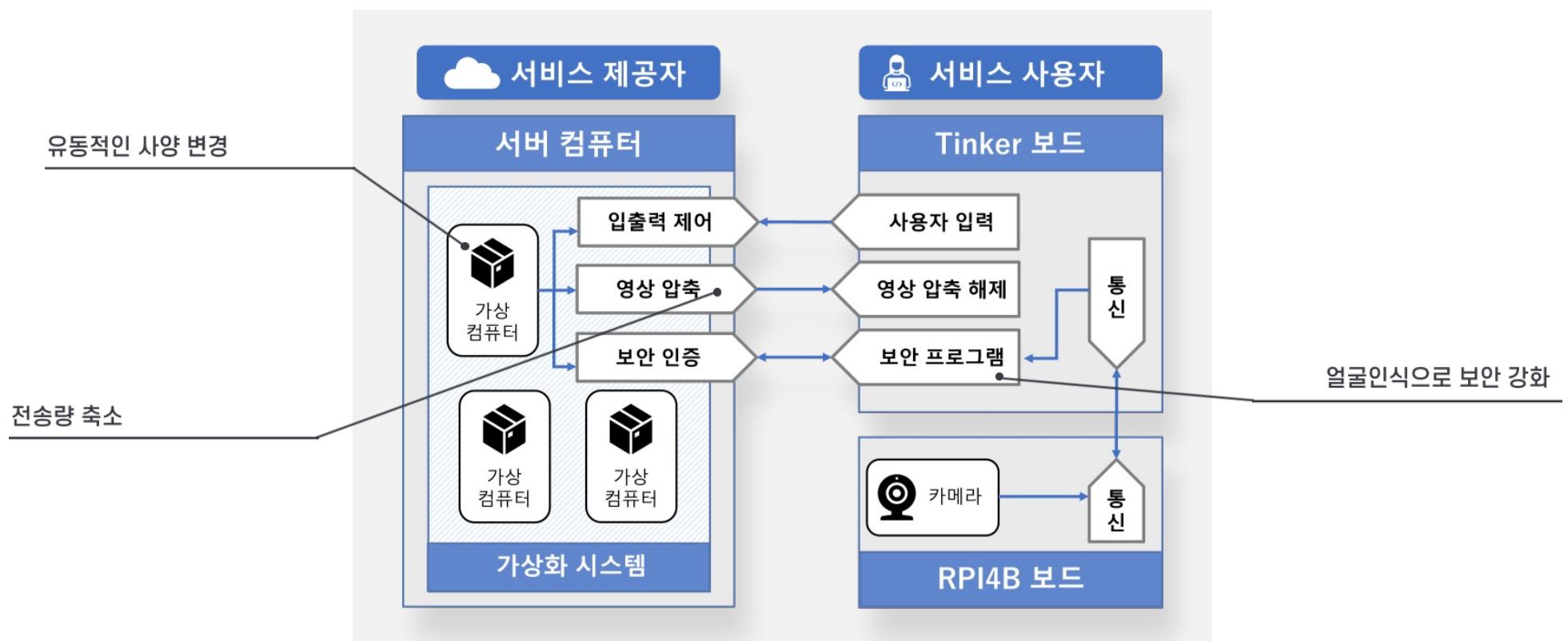
## RFB 프로토콜은 활용하는 VNC

1. VNC는 가장 대중적으로 사용되는 원격 제어 솔루션이며, 대부분 운영체제와 호환됨.
2. VNC는 호스트의 화면을 캡처하여 사용자에게 전송하므로 성능적으로 느림.
3. 또한, 데이터는 H.264 인코딩을 활용하여 전송 용량을 낮춤.
4. VNC는 X11로부터 수신되는 데이터를 기반으로 VNC Client에 전송하는 구조임.
5. 구조적으로는 문제가 많지만,  
별도의 설치 없이 원격제어가 가능하므로  
여전히 많이 사용되어지고 있음.



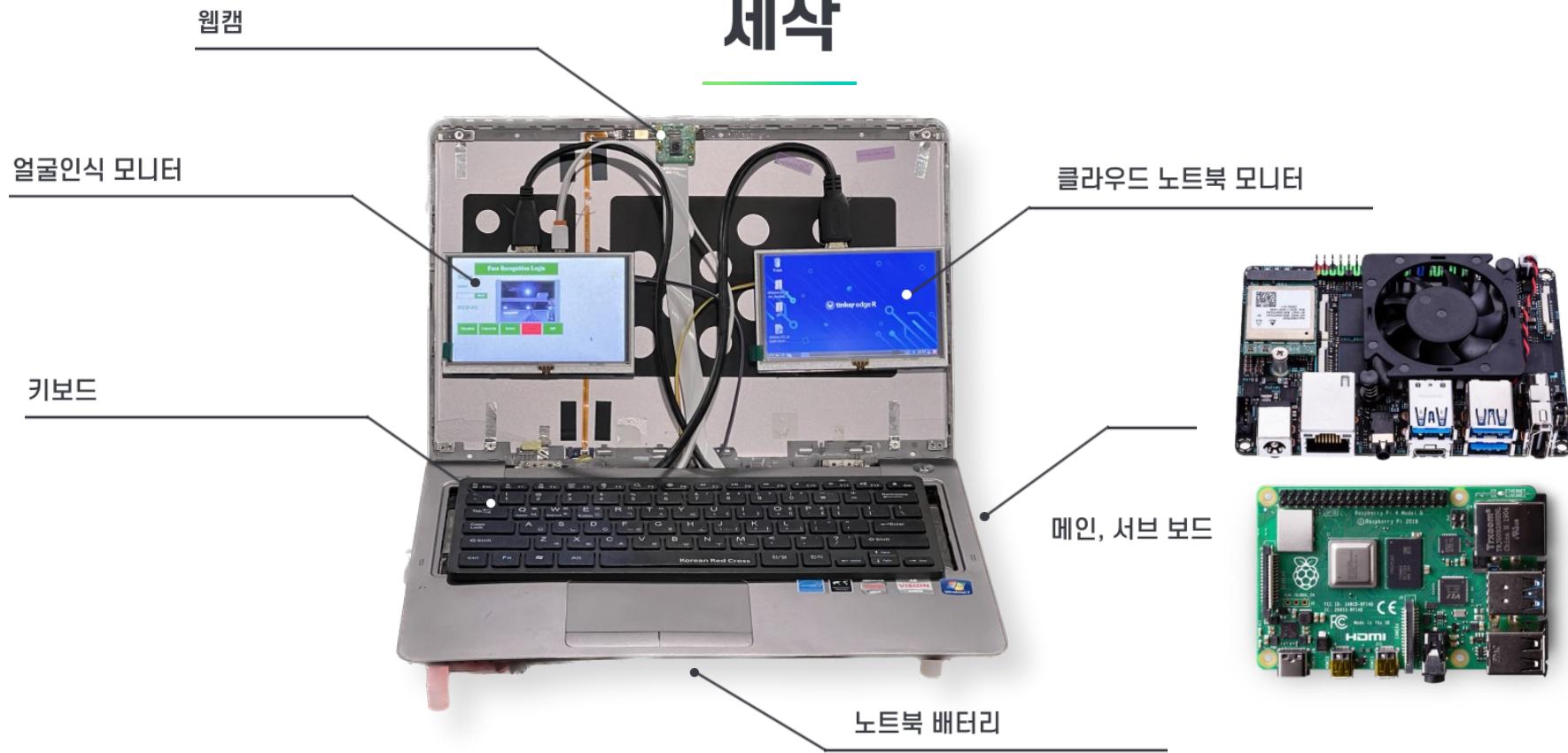
소프트웨어

# 제작



하드웨어

# 제작



# 01 인프라 구축 과정



클라우드 노트북의 사양을  
취향껏 커스터마이징하세요.  
다양한 서비스의 맞춤 클라우드 노트북의 기본 사양을 제공합니다.

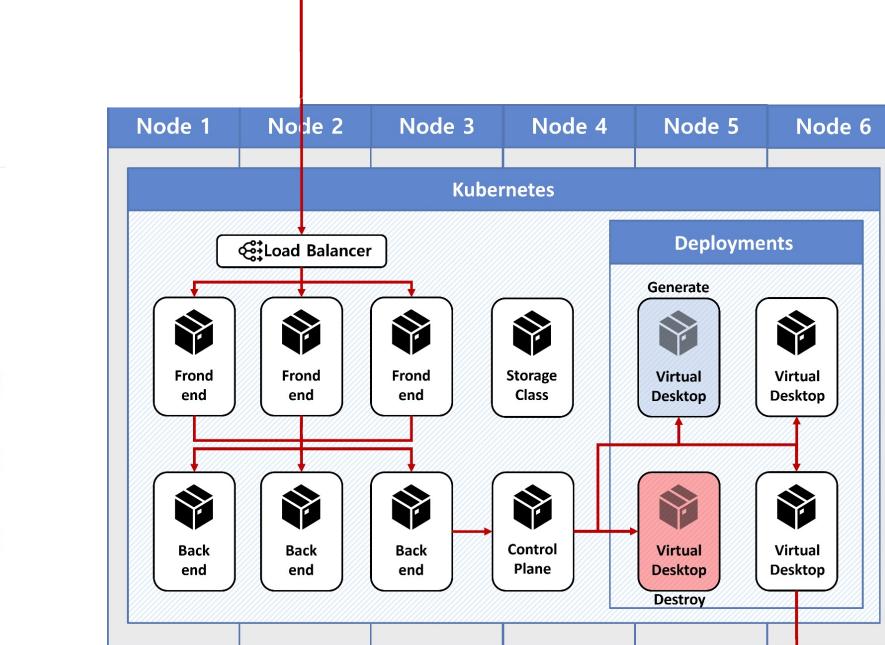
AI 개발 환경이 필요하신가요?

① OS 선택

② CPU/RAM

③ GPU

④ Storage



- 1대의 스토리지 서버와 4대의 쿠버네티스 노드로 구성하였음.
- 이를 통해, 사용자의 VDI 환경은 고가용성을 지님.
- 인프라에 접근하기 위한 전용 하드웨어를 구성하여, 높은 보안성을 제공함.



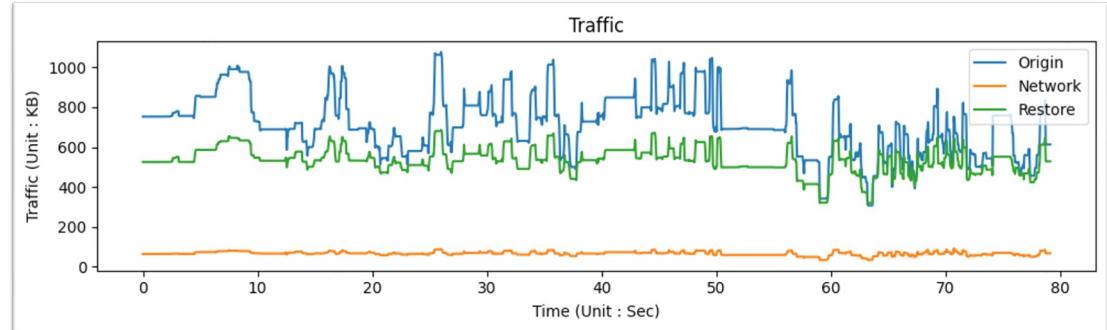
# 05 실험 정의 및 실험 결과

## 실험 정의

- ✓ 실험 정의 :
  - ✓ 제안한 인프라 환경을 활용한 상황에서 트래픽 비용 감소량 확인

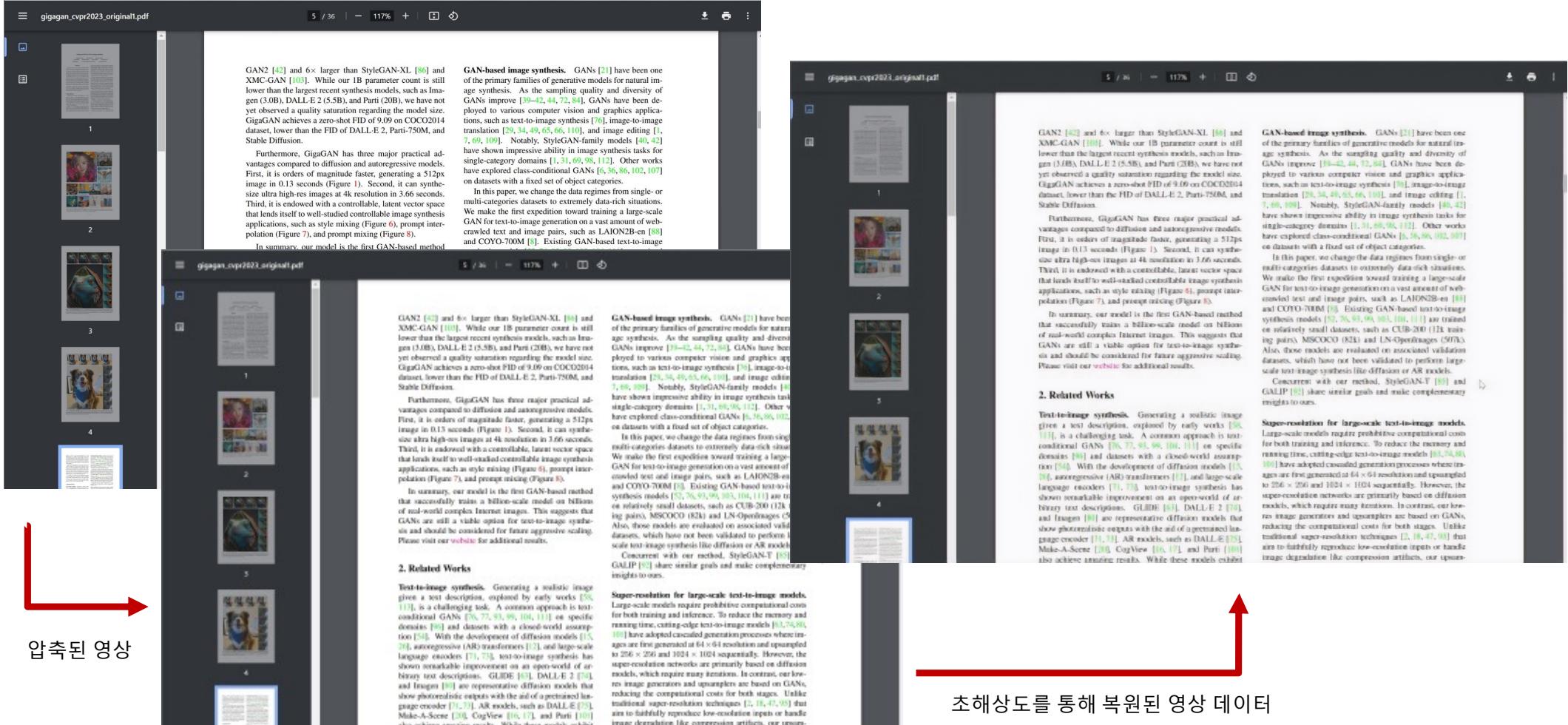
- ✓ 실험 환경 :
  - ✓ VNC와 SRGAN을 사용하는 환경임.
  - ✓ 주로 글자를 읽는 상황이 많은 인터넷 브라우징 환경으로 가정.
  - ✓ 4x SRGAN 기반으로 실험을 수행함.

## 실험 결과



- ✓ 통신비용 :
  - ✓ 기존 VNC 대비 7.2배 통신 비용 절감
- ✓ 추론 성능 :
  - ✓ Tinker EdgeR에 탑재된 NPU에서 4~5 IPS 수행함.
  - ✓ VNC는 평균 2~3 FPS로 동작하였음.
- ✓ 복원 성능 :
  - ✓ 평균적으로 약 75% 복원을 수행함.

# 05 결과



GAN2 [42] and 6x larger than StyleGAN-XL [86] and XMC-GAN [103]. While our 1B parameter count is still lower than the largest recent synthesis models, such as Imagen (3.0B), DALL-E 2 (5.5B), and Parti (20B), we have not yet observed a quality saturation regarding the model size. GigaGAN achieves a zero-shot FID of 9.09 on COCO2014 dataset, lower than the FID of DALL-E 2, Parti-750M, and Stable Diffusion.

Furthermore, GigaGAN has three major practical advantages compared to diffusion and autoregressive models. First, it is orders of magnitude faster, generating a 512px image in 0.13 seconds (Figure 1). Second, it can synthesize ultra high-res images at 4k resolution in 3.66 seconds.

Third, it is endowed with a controllable, latent vector space that lends itself to well-studied controllable image synthesis applications, such as style mixing (Figure 6), prompt interpolation (Figure 7), and prompt mixing (Figure 8).

In summary, our model is the first GAN-based method

**GAN-based image synthesis.** GANs [21] have been one of the primary families of generative models for natural image synthesis. As the sampling quality and diversity of GANs improve [39–42, 44, 72, 84], GANs have been deployed to various computer vision and graphics applications, such as text-to-image synthesis [76], image-to-image translation [29, 34, 49, 65, 66, 110], and image editing [1, 7, 69, 109]. Notably, StyleGAN-family models [40, 42] have shown impressive ability in image synthesis tasks for single-category domains [1, 31, 69, 98, 112]. Other works have explored class-conditional GANs [6, 36, 86, 102, 107]

on datasets with a fixed set of object categories. In this paper, we change the data regimes from single- or multi-categories datasets to extremely data-rich situations. We make the first expedition toward training a large-scale GAN for text-to-image generation on a vast amount of web-crawled text and image pairs, such as LAION2B-en [88] and COCO-700M [3]. Existing GAN-based text-to-image synthesis models [37, 76, 83, 96, 103, 104, 113] are trained on relatively small datasets, such as CUB-200 (11k training pairs), MSCOCO (83k) and LN-OpenImages (597k). Also, those models are evaluated on associated validation datasets, which have not been validated to perform large-scale text-to-image synthesis like diffusion or AIR models.

Furthermore, GigaGAN has three major practical advantages compared to diffusion and autoregressive models. First, it is orders of magnitude faster, generating a 512px image in 0.13 seconds (Figure 1). Second, it can synthesize ultra high-res images at 4k resolution in 3.66 seconds.

Third, it is endowed with a controllable, latent vector space that lends itself to well-studied controllable image synthesis applications, such as style mixing (Figure 6), prompt interpolation (Figure 7), and prompt mixing (Figure 8).

In summary, our model is the first GAN-based method that successfully trains a billion-scale model on billions of real-world complex Internet images. This suggests that GANs are still a viable option for text-to-image synthesis and should be considered for future aggressive scaling. Please visit our [website](#) for additional results.

## 2. Related Works

**Text-to-image synthesis.** Generating a realistic image given a text description, explored by early works [58, 113], is a challenging task. A common approach is text-conditional GANs [26, 77, 93, 99, 104, 111] on specific domains [98] and datasets with a closed-world assumption [54]. With the development of diffusion models [115, 34], autoregressive (AIR) transformers [17], and large-scale language encoders [73, 75], text-to-image synthesis has shown remarkable improvement on an open-world of arbitrary text descriptions. GLIDE [16], DALL-E 2 [78], and Imagen [109] are representative diffusion models that show photorealistic outputs with the aid of a pretrained language encoder [73, 75]. AIR models, such as DALL-E [75], Make-A-Scene [30], CogView [16, 17], and Parti [100] also achieve amazing results. While these models exhibit

GAN2 [42] and 6x larger than StyleGAN-XL [86] and XMC-GAN [103]. While our 1B parameter count is still lower than the largest recent synthesis models, such as Imagen (3.0B), DALL-E 2 (5.5B), and Parti (20B), we have not yet observed a quality saturation regarding the model size. GigaGAN achieves a zero-shot FID of 9.09 on COCO2014 dataset, lower than the FID of DALL-E 2, Parti-750M, and Stable Diffusion.

Furthermore, GigaGAN has three major practical advantages compared to diffusion and autoregressive models. First, it is orders of magnitude faster, generating a 512px image in 0.13 seconds (Figure 1). Second, it can synthesize ultra high-res images at 4k resolution in 3.66 seconds.

Third, it is endowed with a controllable, latent vector space that lends itself to well-studied controllable image synthesis applications, such as style mixing (Figure 6), prompt interpolation (Figure 7), and prompt mixing (Figure 8).

In this paper, we change the data regimes from single- or multi-categories datasets to extremely data-rich situations. We make the first expedition toward training a large-scale GAN for text-to-image generation on a vast amount of web-crawled text and image pairs, such as LAION2B-en [88] and COCO-700M [3]. Existing GAN-based text-to-image synthesis models [37, 76, 83, 96, 103, 104, 113] are trained on relatively small datasets, such as CUB-200 (11k training pairs), MSCOCO (83k) and LN-OpenImages (597k). Also, those models are evaluated on associated validation datasets, which have not been validated to perform large-scale text-to-image synthesis like diffusion or AIR models.

Concurrent with our method, StyleGAN-T [85] and GALIP [102] share similar goals and make complementary insights to ours.

**GAN-based image synthesis.** GANs [21] have been one of the primary families of generative models for natural image synthesis. As the sampling quality and diversity of GANs improve [39–42, 44, 72, 84], GANs have been deployed to various computer vision and graphics applications, such as text-to-image synthesis [76], image-to-image translation [29, 34, 49, 65, 66, 110], and image editing [1, 7, 69, 109]. Notably, StyleGAN-family models [40, 42] have shown impressive ability in image synthesis tasks for single-category domains [1, 31, 69, 98, 112]. Other works have explored class-conditional GANs [6, 36, 86, 102, 107]

on datasets with a fixed set of object categories. In this paper, we change the data regimes from single- or multi-categories datasets to extremely data-rich situations. We make the first expedition toward training a large-scale GAN for text-to-image generation on a vast amount of web-crawled text and image pairs, such as LAION2B-en [88] and COCO-700M [3]. Existing GAN-based text-to-image synthesis models [37, 76, 83, 96, 103, 104, 113] are trained on relatively small datasets, such as CUB-200 (11k training pairs), MSCOCO (83k) and LN-OpenImages (597k). Also, those models are evaluated on associated validation datasets, which have not been validated to perform large-scale text-to-image synthesis like diffusion or AIR models.

In this paper, our model is the first GAN-based method that successfully trains a billion-scale model on billions of real-world complex Internet images. This suggests that GANs are still a viable option for text-to-image synthesis and should be considered for future aggressive scaling. Please visit our [website](#) for additional results.

## 2. Related Works

**Text-to-image synthesis.** Generating a realistic image given a text description, explored by early works [58, 113], is a challenging task. A common approach is text-conditional GANs [26, 77, 93, 99, 104, 111] on specific domains [98] and datasets with a closed-world assumption [54]. With the development of diffusion models [115, 34], autoregressive (AIR) transformers [17], and large-scale language encoders [73, 75], text-to-image synthesis has shown remarkable improvement on an open-world of arbitrary text descriptions. GLIDE [16], DALL-E 2 [78], and Imagen [109] are representative diffusion models that show photorealistic outputs with the aid of a pretrained language encoder [73, 75]. AIR models, such as DALL-E [75], Make-A-Scene [30], CogView [16, 17], and Parti [100] also achieve amazing results. While these models exhibit

GAN2 [42] and 6x larger than StyleGAN-XL [86] and XMC-GAN [103]. While our 1B parameter count is still lower than the largest recent synthesis models, such as Imagen (3.0B), DALL-E 2 (5.5B), and Parti (20B), we have not yet observed a quality saturation regarding the model size. GigaGAN achieves a zero-shot FID of 9.09 on COCO2014 dataset, lower than the FID of DALL-E 2, Parti-750M, and Stable Diffusion.

Furthermore, GigaGAN has three major practical advantages compared to diffusion and autoregressive models. First, it is orders of magnitude faster, generating a 512px image in 0.13 seconds (Figure 1). Second, it can synthesize ultra high-res images at 4k resolution in 3.66 seconds.

Third, it is endowed with a controllable, latent vector space that lends itself to well-studied controllable image synthesis applications, such as style mixing (Figure 6), prompt interpolation (Figure 7), and prompt mixing (Figure 8).

In this paper, our model is the first GAN-based method that successfully trains a billion-scale model on billions of real-world complex Internet images. This suggests that GANs are still a viable option for text-to-image synthesis and should be considered for future aggressive scaling. Please visit our [website](#) for additional results.

Concurrent with our method, StyleGAN-T [85] and GALIP [102] share similar goals and make complementary insights to ours.

**Super-resolution for large-scale text-to-image models.** Large-scale models require prohibitive computational costs for both training and inference. To reduce the memory and running time, cutting-edge text-to-image models [63, 74, 80, 100] have adopted cascaded generation processes where images are first generated at 64 × 64 resolution and upsampled to 256 × 256 and 1024 × 1024 sequentially. However, the super-resolution networks are primarily based on diffusion models, which require many iterations. In contrast, our low-res image generators and upscalers are based on GANs, reducing the computational costs for both stages. Unlike traditional super-resolution techniques [2, 18, 47, 93] that aim to faithfully reproduce low-resolution inputs or handle image degradation like compression artifacts, our upscalers

## 05 결과

