SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Rethinking Storage for the AI/ML Era

Disaggregated Powered with FDP

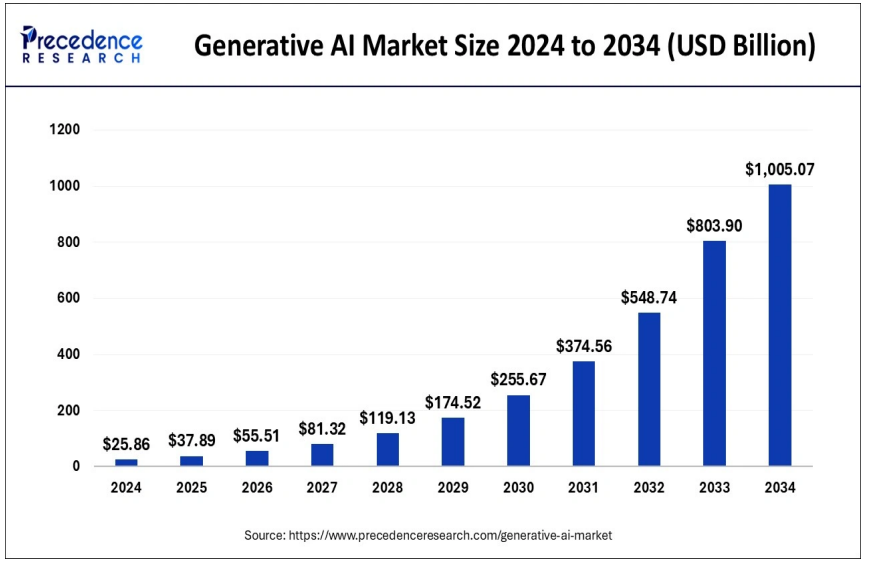
Sathish Kumar M (Associate Tech Director), Samsung Semiconductor Inc
Amit Devgan (Senior Staff Engg), Samsung Semiconductor Inc
Arun Kumar Singh (Senior Staff Engg), Samsung Semiconductor Inc
Ratish Gopinath (Associate Staff Engg), Samsung Semiconductor Inc

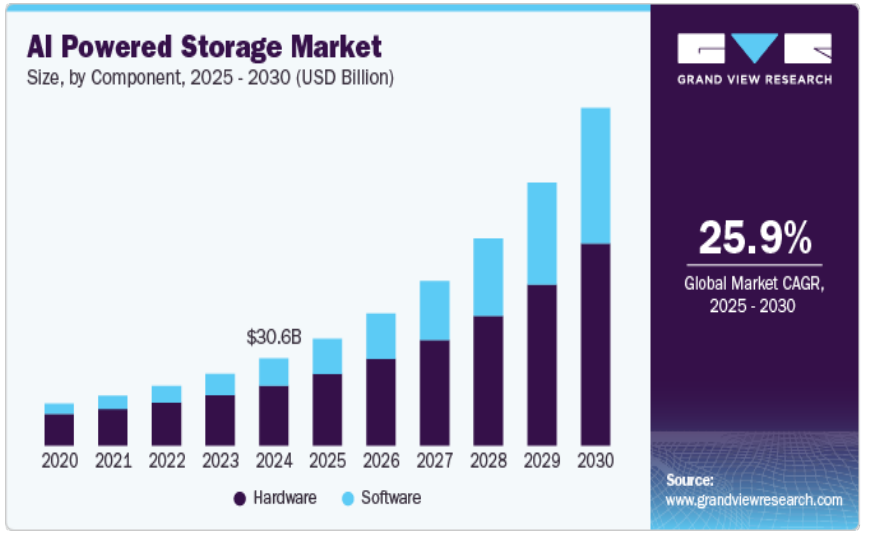
Agenda

- Overview Generative AI
- Understanding of Stable Diffusion(SD)
- Storage Aspects of SD
- On-Prem Deployment with NVMe-oF
- Ecosystem and workload
- Storage Challenges and Characteristics
- FDP and WAF
- Summary

Generative AI

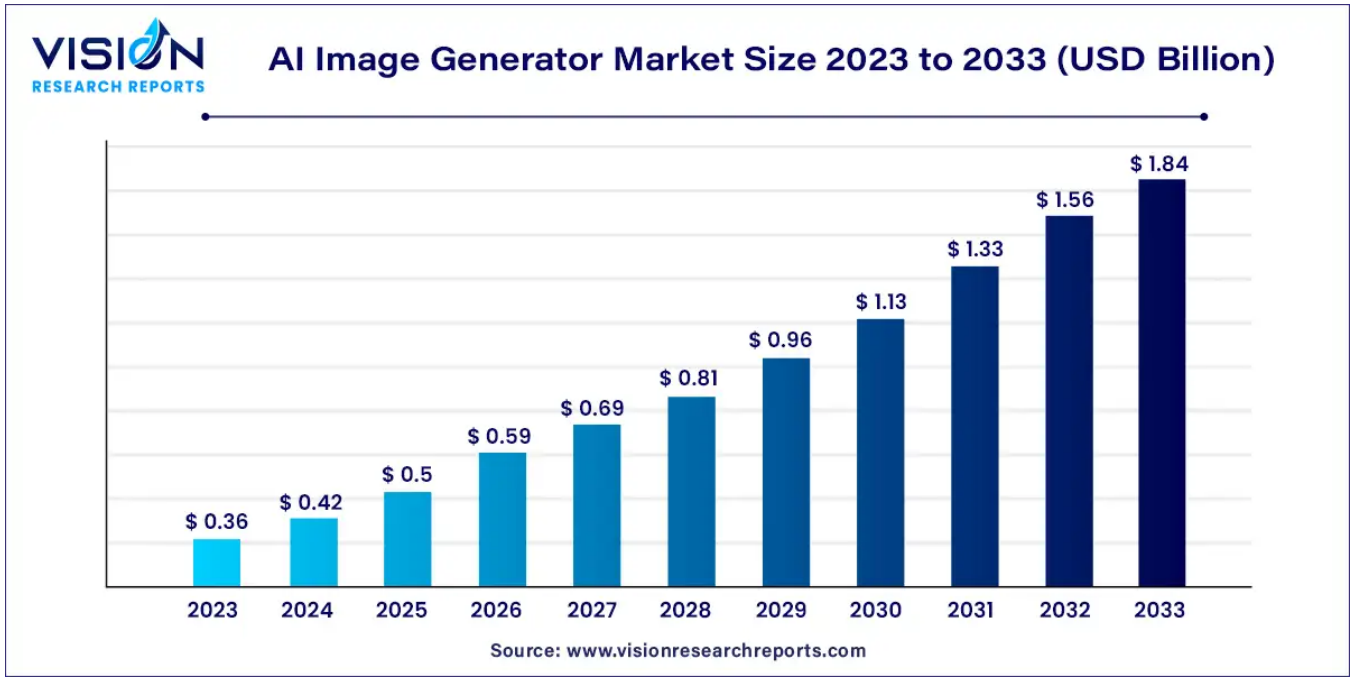
- AI models that can create new content
  - Images, Text, Audio, Video
- Revolutionizing various industries
  - Manufacturing, design, gaming, education, entertainment, and research
- GenAI increases Storage demand
  - Training, Checkpointing, Inference, RAG
  - Performance, Reliable, Capacity & Secure





Stable Diffusion

- **Deep learning, text-to-image model released in 2022**
  - **Open source Gen AI technology from Stability AI**
- **Use cases: Design & Art, Entertainment, Gaming, Marketing, Healthcare, Fashion**

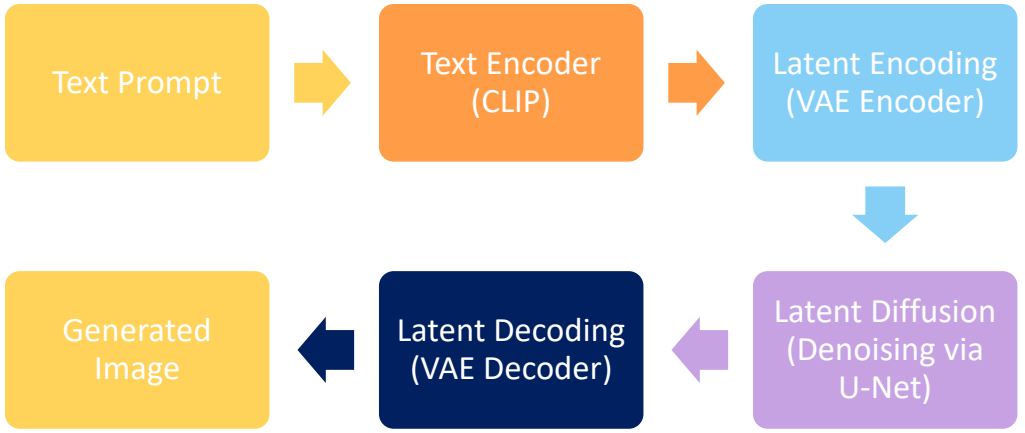


An image generated with Stable Diffusion
Text prompt: a photograph of an astronaut riding a horse

Understanding Stable Diffusion Model

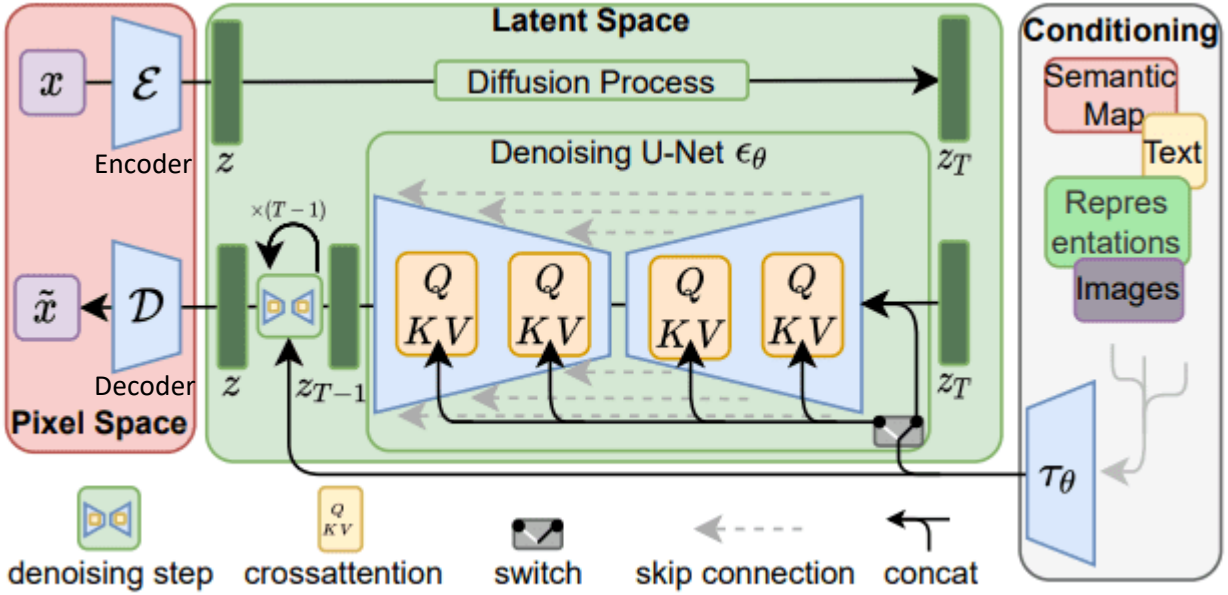
- Open-source
  - Runs on PyTorch and HuggingFace open-source frameworks

- Core components:
  - CLIP(Contrastive Language-Image retraining):
    - Text encoding
  - U-Net
    - Denoising
  - Variational Auto Encoder(VAE)
    - Image encoding/decoding



LDM(Latent Diffusion Model) Architecture

- Pixel space: Encoder
  - Image to pixel(x)
  - Compress to latent image (x → z)
- Latent Space
  - Diffusion Process:
    - Add noises gradually(z → z_T)
  - Denoising U-Net
    - Remove noise steps (z_T → z)
    - With guidance (self/cross attention)
- Conditioning
  - Semantic Map, Text, Reference images
  - Inputs to Denoising process make better output latent
- Pixel Space: Decoder
  - Latent image(z) to high resolution image (ẋ)



Source: https://arxiv.org/pdf/2112.10752

Storage Components in Stable Diffusion

- Pre-trained models
- Intermediate latent data (temp noised images)
- Output images store
- Prompts and Logs

Storage-Intensive Components

Model Load

- Read
- Stable Diffusion 3.5 Large: ~20GB (fp16), 8.1 billion parameters.
- Stable Diffusion 3.5 Medium: ~10GB (fp16), 2.5 billion parameters.
- SDXL Base Model: ~6.6GB.
- Checkpoint models: 2-7GB

Generated Images

- Write-heavy
- 5–50 MB/image

Goal

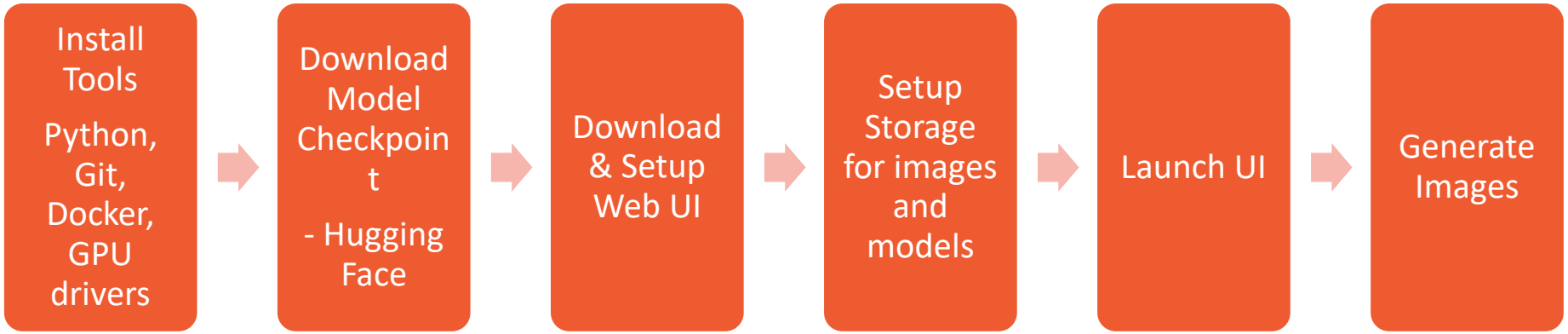
- On-Prem deployment
- Scalable storage for SD deployment
- Stress the SD workload on storage
- Analyze and understand IO during SD
- Increase the lifespan of storage device

Cloud vs On-Prem Stable Diffusion

Deployment	Cloud	On-Prem
Privacy	No	Yes
Control and Customization	Limited	Full
Cost Effective	Short Term	Long Term
Access	API rate	Unlimited

Deployment of Stable Diffusion

- Install Tools
- Clone pre-trained model
  - https://huggingface.co/CompVision/stable-diffusion
- Setup inference environment
- Provision Storage for Model and Images
- Generate Images from prompts



H/W Setup

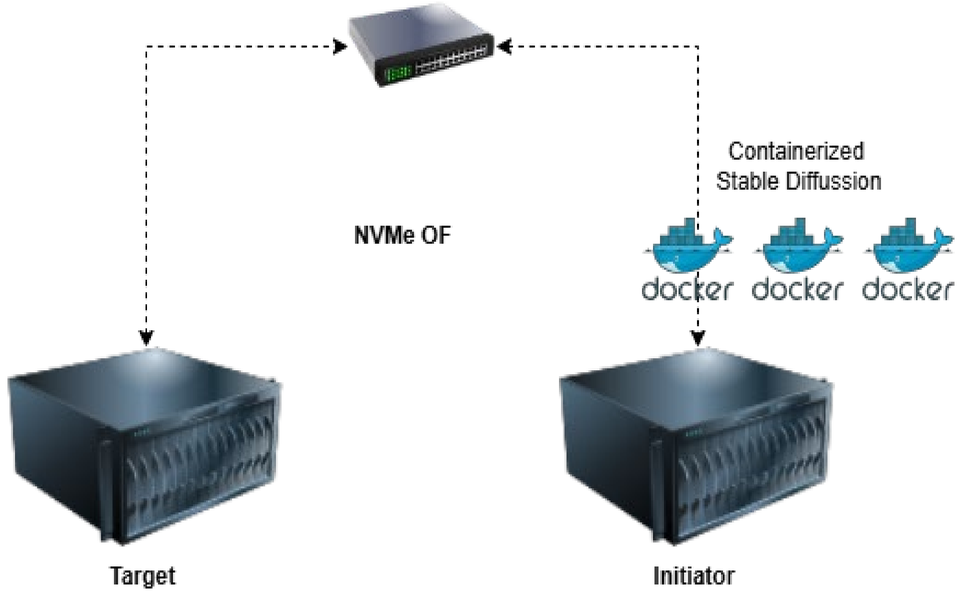
Stable Diffusion Service
Intel x86 server
H100 80GB PCIe GPU
1 TB DDR5
100 GBE
CUDA version: 13.0

100G Network

Target
AMD Genoa Storage Server
32 x 16 TB TLC NVMe drives
128 GB DDR5
100 GBE
SPDK 25.05 (NVMe oF/TCP)

- Flexibility to disaggregate compute and storage
- As user grows, compute and storage expanded separately
- Stable diffusion is leveraging the backend storage.
  - Model
  - Generated Images
  - Checkpoints

Containerization of Stable Diffusion



CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS	NAMES
d6b4c51af179	sd-auto:78	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7925->7925/tcp, [::]:7925->7925/tcp	stable-diffusion-webui-docker_25-auto-1
fc826bb7718f	ee697c1de72a	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7924->7924/tcp, [::]:7924->7924/tcp	stable-diffusion-webui-docker_24-auto-1
6796bf8da9d2	0c2a5a9286bc	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7923->7923/tcp, [::]:7923->7923/tcp	stable-diffusion-webui-docker_23-auto-1
b82a78d382b1	a3691a222cd8	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7922->7922/tcp, [::]:7922->7922/tcp	stable-diffusion-webui-docker_22-auto-1
4a604795f238	b4a5431a1d88	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7921->7921/tcp, [::]:7921->7921/tcp	stable-diffusion-webui-docker_21-auto-1
fa702017b466	a82393e82017	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7920->7920/tcp, [::]:7920->7920/tcp	stable-diffusion-webui-docker_20-auto-1
9cb9318d37cc	0c1734b49694	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7919->7919/tcp, [::]:7919->7919/tcp	stable-diffusion-webui-docker_19-auto-1
fa716833ae57	2b8aa81b34d8	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7918->7918/tcp, [::]:7918->7918/tcp	stable-diffusion-webui-docker_18-auto-1
b6cd402c23de	c189af4ba388	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7917->7917/tcp, [::]:7917->7917/tcp	stable-diffusion-webui-docker_17-auto-1
61a56720bec4	524436b3d3ca	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7916->7916/tcp, [::]:7916->7916/tcp	stable-diffusion-webui-docker_16-auto-1
73d5ce762fb2	4f2db3a2ae31	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7915->7915/tcp, [::]:7915->7915/tcp	stable-diffusion-webui-docker_15-auto-1
28c0fe1f41c0	1ab793ddb45a	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7914->7914/tcp, [::]:7914->7914/tcp	stable-diffusion-webui-docker_14-auto-1
6feb2ed0cf3d	f65e9711f50b	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7913->7913/tcp, [::]:7913->7913/tcp	stable-diffusion-webui-docker_13-auto-1
3bbef0e837c8	267ce1db2084	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7912->7912/tcp, [::]:7912->7912/tcp	stable-diffusion-webui-docker_12-auto-1
4aa766fdd5eb	d49e5c3bffc6	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7911->7911/tcp, [::]:7911->7911/tcp	stable-diffusion-webui-docker_11-auto-1
c5fb5380dc4f	560bacaf3d59	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7910->7910/tcp, [::]:7910->7910/tcp	stable-diffusion-webui-docker_10-auto-1
01f3e79c7c43	bcba0c917266	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7909->7909/tcp, [::]:7909->7909/tcp	stable-diffusion-webui-docker_09-auto-1
db4ff0e3be8f	681b421a88fd	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7908->7908/tcp, [::]:7908->7908/tcp	stable-diffusion-webui-docker_08-auto-1
0eb8473c2d73	d384ad37d5d1	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7907->7907/tcp, [::]:7907->7907/tcp	stable-diffusion-webui-docker_07-auto-1
8c8f0ee6c0af	2a0be625c583	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7906->7906/tcp, [::]:7906->7906/tcp	stable-diffusion-webui-docker_06-auto-1
d3848140dc36	d5dee10bced0	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7905->7905/tcp, [::]:7905->7905/tcp	stable-diffusion-webui-docker_05-auto-1
0b81773cbe75	c7e5c7dc785c	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7904->7904/tcp, [::]:7904->7904/tcp	stable-diffusion-webui-docker_04-auto-1
9f24278ef41d	8904d7039db9	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7903->7903/tcp, [::]:7903->7903/tcp	stable-diffusion-webui-docker_03-auto-1
11bc8b0a81fd	a44ea680b0c9	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7901->7901/tcp, [::]:7901->7901/tcp	stable-diffusion-webui-docker_01-auto-1
c348da613a77	a44ea680b0c9	"/docker/entrypoint.…"	5 days ago	Up 5 days	7860/tcp, 0.0.0.0:7902->7902/tcp, [::]:7902->7902/tcp	stable-diffusion-webui-docker_02-auto-1

Muti-container status

- Single SD instance – Serial processing
- On-Prem multiuser environment requires parallel processing.
- Leveraging the Docker containerization for parallelism.
  - Load model and generates images
- Scale independently as user grows

Stable Diffusion Workload Generator

- Load generator script
  - Multithreaded, RestAPI using 1000's of random user prompts for image generation
  - Runs on 25 containers

workload.py

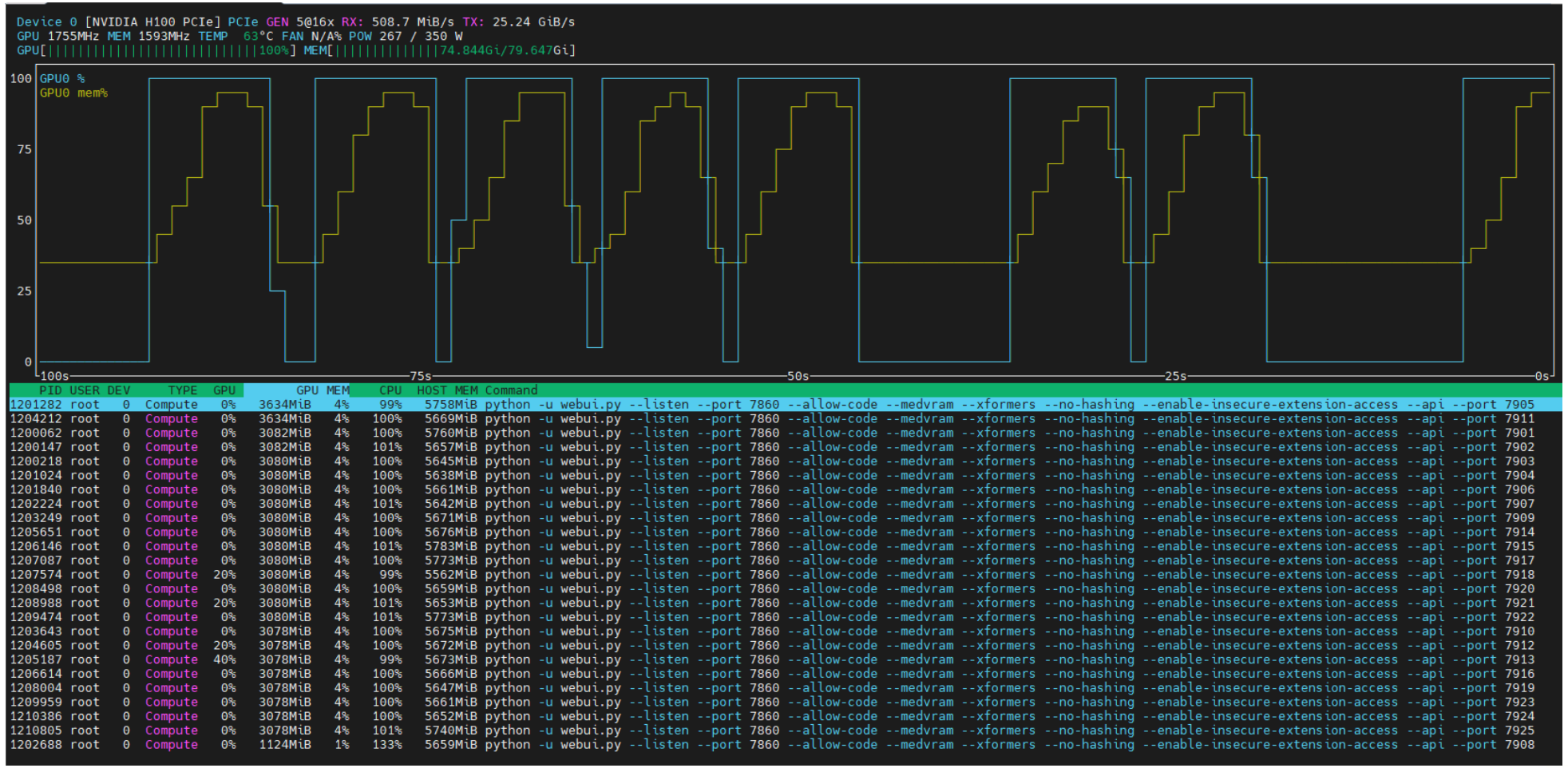
```python
import requests
import random
import threading

# Define the URL and the payload to send.
server = "10.10.40.138"
ports = range(7901,7926)

#Generate random prompt
with open("prompt.source",'r') as f:
        prompts = f.readlines()

#Run Stable diffusion
def stable_diffusion(name, prompt, server, port):
        url = "http://"+server+ ":" + str(port)
        payload = {
            "prompt": prompt,
            "steps": 5,
            'save_images': True
        }

        # Send said payload to said URL through the API.
        response = requests.post(url=f'{url}/sdapi/v1/txt2img', json=payload)
        r = response.json()

#Run Stable diffusion
while(True):
        threads = []
        for port in ports:
                random_prompt = random.choice(prompts).replace("\n",'')
                # Create multiple threads running the same function with different arguments
                thread = threading.Thread(target=stable_diffusion, args=(port, random_prompt, server,port))
                threads.append(thread)
                thread.start()

        for thread in threads:
                thread.join()
```

prompt.source

```
Portraits & People
A hyper-realistic portrait of an old sailor with deep wrinkles, wearing
A futuristic cyberpunk woman with glowing tattoos, neon city background
A royal Indian maharaja portrait in traditional attire, oil painting st
A smiling African woman wearing a colorful head wrap, Canon EOS photo r
A Victorian gentleman with a monocle, sepia-toned vintage photograph.
A close-up portrait of a Viking warrior with braided hair, snow falling
A Japanese geisha in a kimono, standing under cherry blossoms, watercol
A superhero costume concept, cinematic concept art.
A medieval knight in shining armor, HDR photography style.
A steampunk inventor woman with brass goggles, workshop background.
Landscapes & Nature
A serene lake at sunrise with mist, photorealism.
A lush jungle with glowing plants, fantasy illustration.
A desert landscape with futuristic wind turbines, concept art.
A snowy mountain peak with aurora borealis, wide-angle shot.
A tropical beach with crystal clear water, drone photography.
A canyon lit by golden hour light, ultra-realistic.
A mystical enchanted forest with fireflies, fantasy digital art.
A volcanic eruption at night, cinematic photography.
A futuristic floating island in the clouds, concept art.
A desert oasis with palm trees, watercolor painting.
3. Fantasy & Mythology
A dragon flying over a medieval castle, fantasy illustration.
A Greek goddess standing in a marble temple, oil painting.
A dark wizard casting a spell, glowing staff, digital art.
A phoenix rising from flames, high detail fantasy art.
A mermaid underwater near glowing corals, dreamy atmosphere.
A Norse god wielding a giant hammer, epic concept art.
A fairy queen with glowing wings in a magical forest.
A dwarven blacksmith forging a sword in his workshop.
A samurai fighting a demon, anime style.
A medieval tavern full of adventurers, D&D fantasy illustration.
4. Sci-Fi & Futuristic
A neon cyberpunk Tokyo street at night, cinematic shot.
A futuristic astronaut exploring an alien desert, concept art.
A massive space station orbiting Earth, ultra detail.
A robot repairing another robot in a workshop, Pixar style.
A dystopian city skyline with flying cars, Blade Runner vibes.
A holographic library in the year 2500, digital art.
A mecha robot standing in rain, anime concept art.
A futuristic soldier in exosuit, battlefield sci-fi concept.
A time machine portal opening in a lab.
A spaceship docking inside a massive alien mothership.
5. Architecture & Interiors
A futuristic glass skyscraper in New York, ultra-realistic.
A medieval castle on a cliff, matte painting style.
```


NVTOP

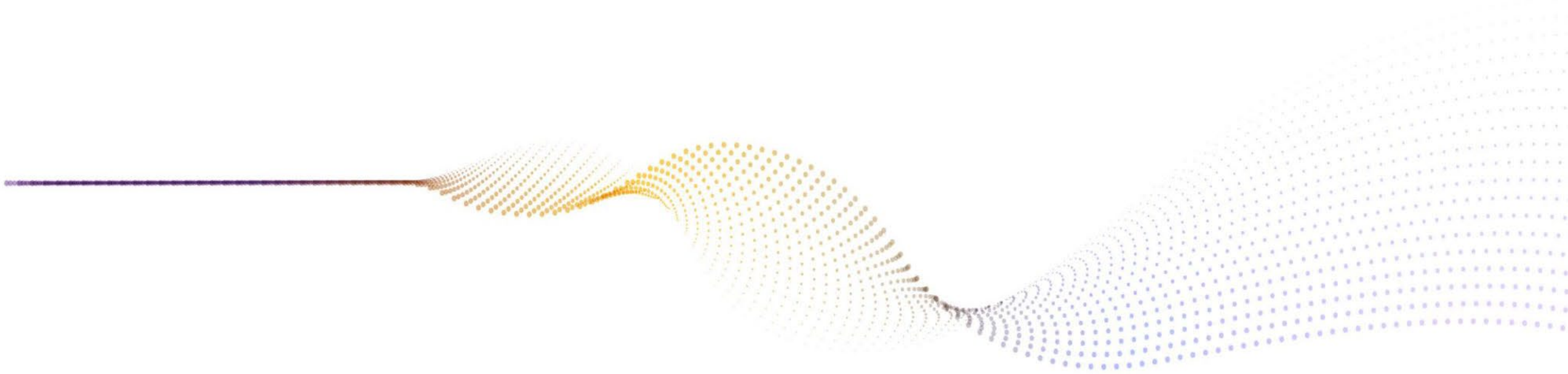
- Stable Diffusion workload stress
  - Reached max utilization of GPU
  - Parallel image generation



Stable Diffusion: Multi-container Image Generation

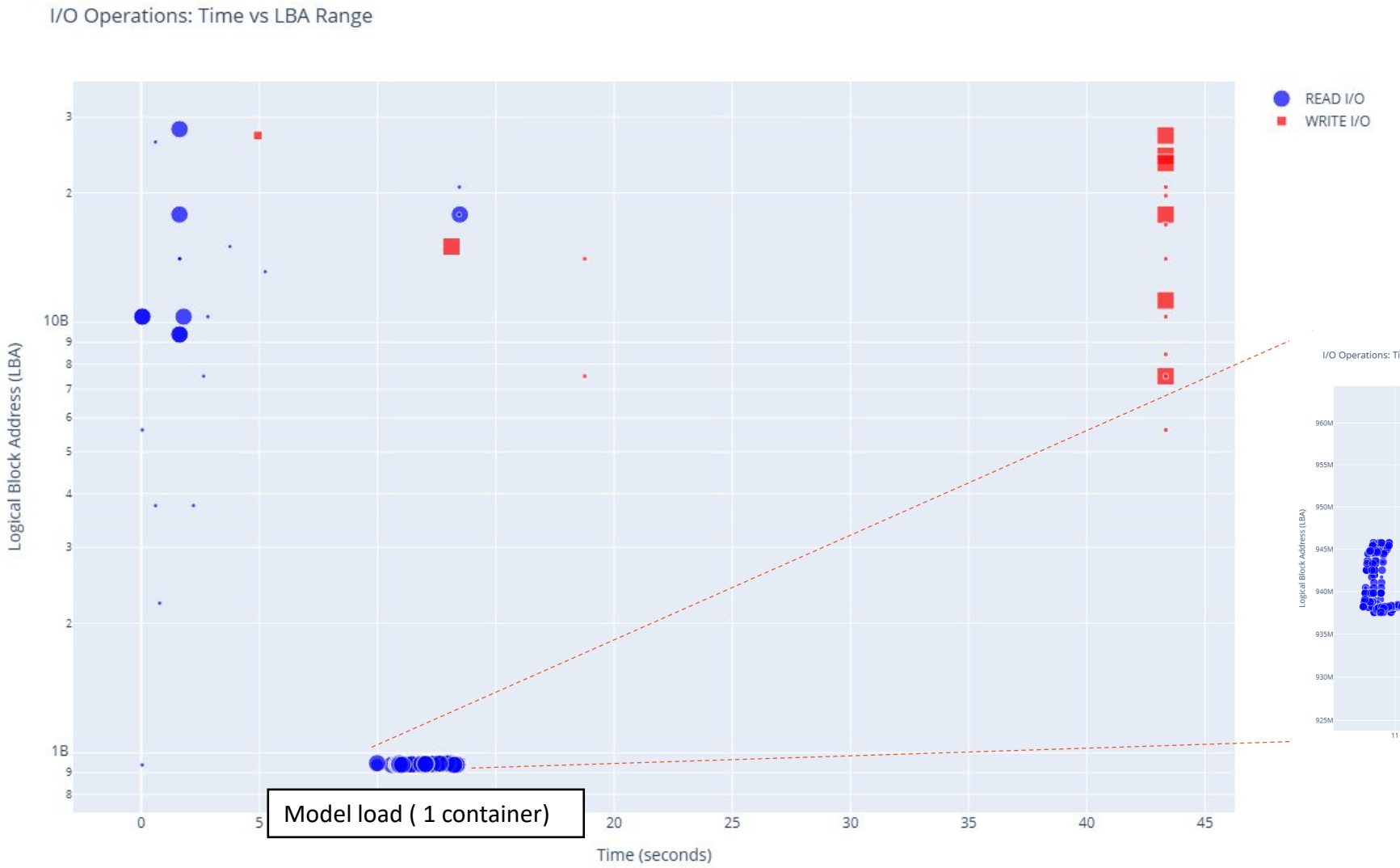
```
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00079-2323040410.png
1.9M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00080-230415262.png
1.9M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00081-230415263.png
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00082-4168479516.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00083-4168479517.png
1.7M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00084-3924127982.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00085-3924127983.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00088-285206579.png
1.3M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00089-285206580.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00092-4044973334.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00093-4044973335.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00094-2982817126.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00095-2982817127.png
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00096-636626183.png
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00097-636626184.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00098-1088849160.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00099-1088849161.png
1.1M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00100-3592828560.png
1.3M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00101-3592828561.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00102-1703555100.png
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00103-1703555101.png
1.5M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00106-1629062207.png
1.6M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00107-1629062208.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00108-2675427530.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00109-2675427531.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00114-3045984609.png
1.1M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00115-3045984610.png
1.3M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00116-2834661385.png
1.6M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00117-2834661386.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00118-2082506908.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00119-2082506909.png
1.9M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00120-3702410395.png
1.9M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00121-3702410396.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00122-4144066135.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00123-4144066136.png
1.3M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00124-3868044023.png
1.4M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00125-3868044024.png
1.3M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00128-1116917025.png
1.2M    /mnt/stablediffusion/stable-diffusion-webui-docker/output/txt2img/2025-08-18/00129-1116917026.png
gpu_server#
```

- Generated images
  - 4GB model
  - 1024 x 1024
  - Parallel Images



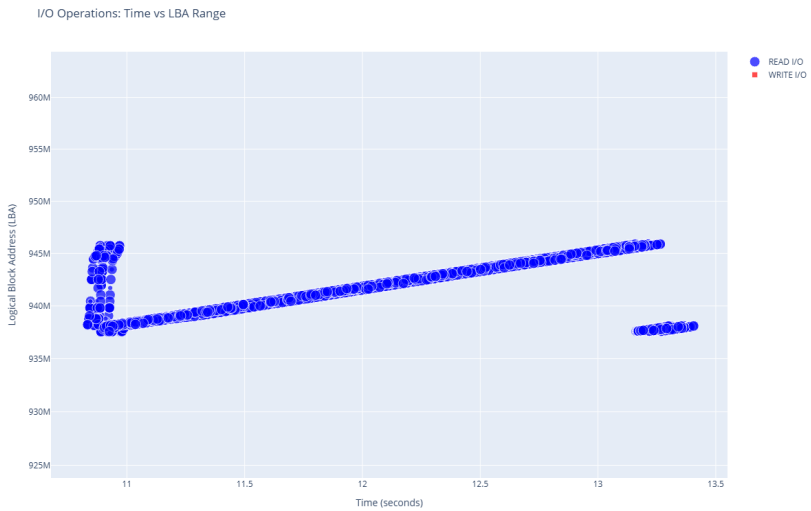
Storage Characteristics

Analysis for Stable Diffusion: Model Load – Single Container

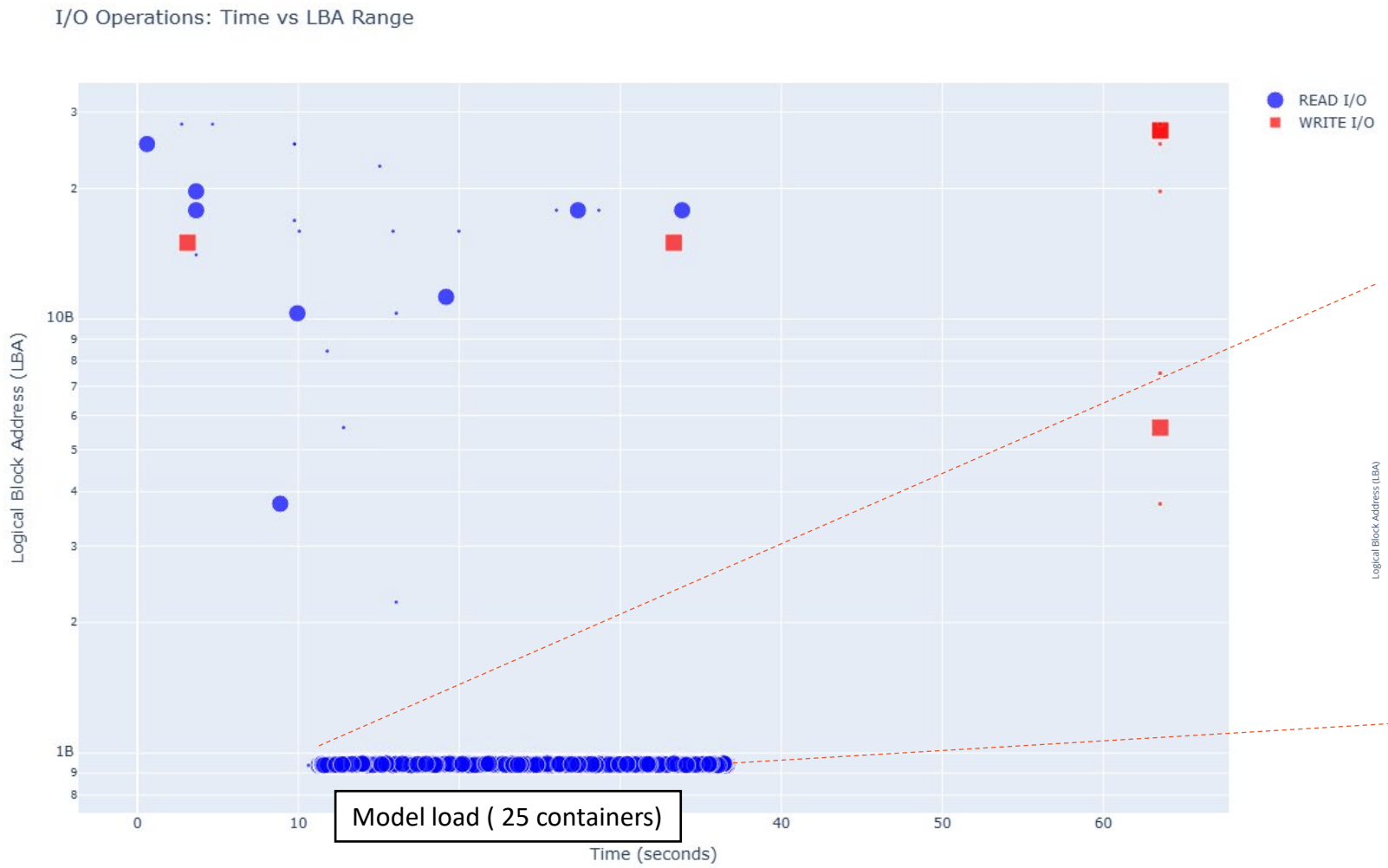


Single Container:

- Sequential Read IO
- LBA is accessed sequentially
- Load time: 2 sec (4GB)

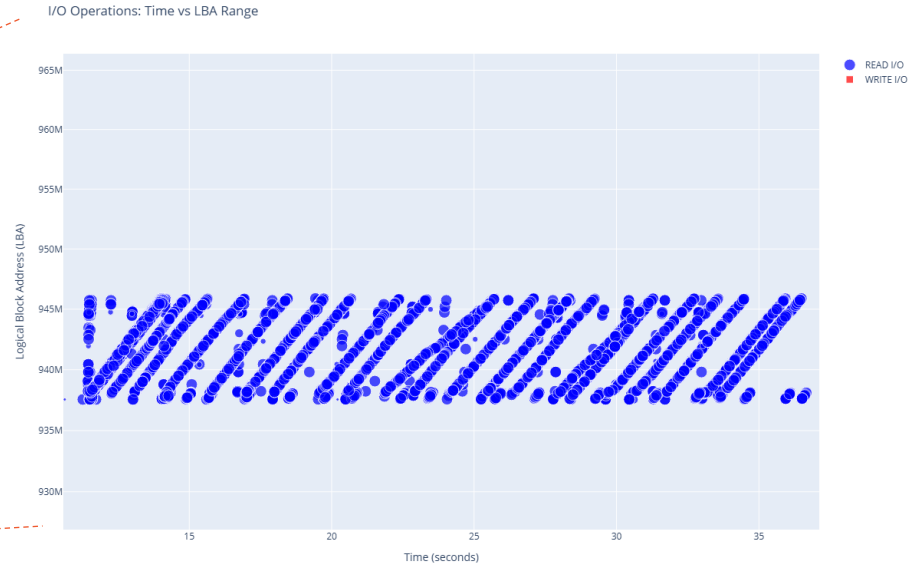


Analysis for Stable Diffusion: Model Load – Multi Container



Multi Container

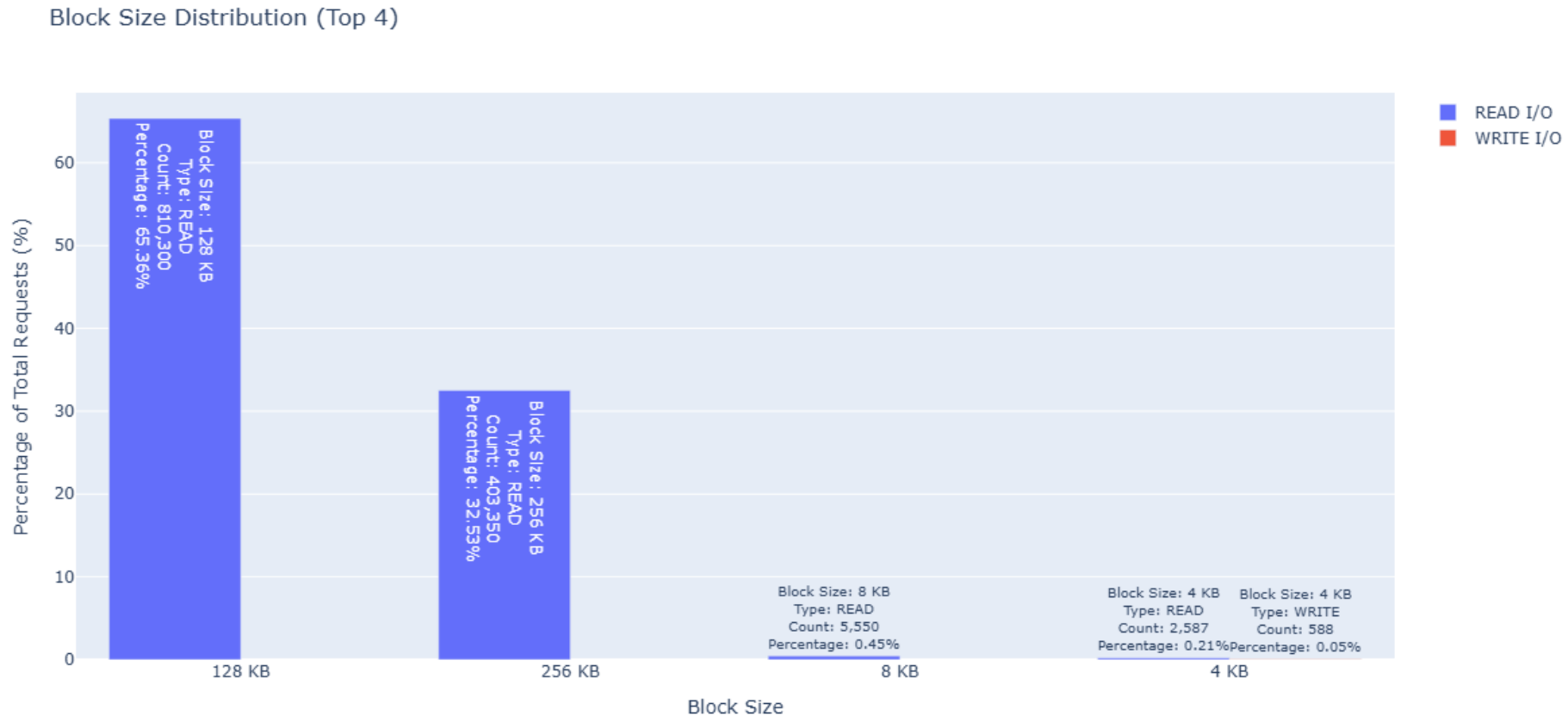
- Random Read IO
- 25 parallel access that forms a Random IO
- Load time: 28 secs (4 GB)



Block Distribution – Model Load

The graph depicts the model load workload on the device

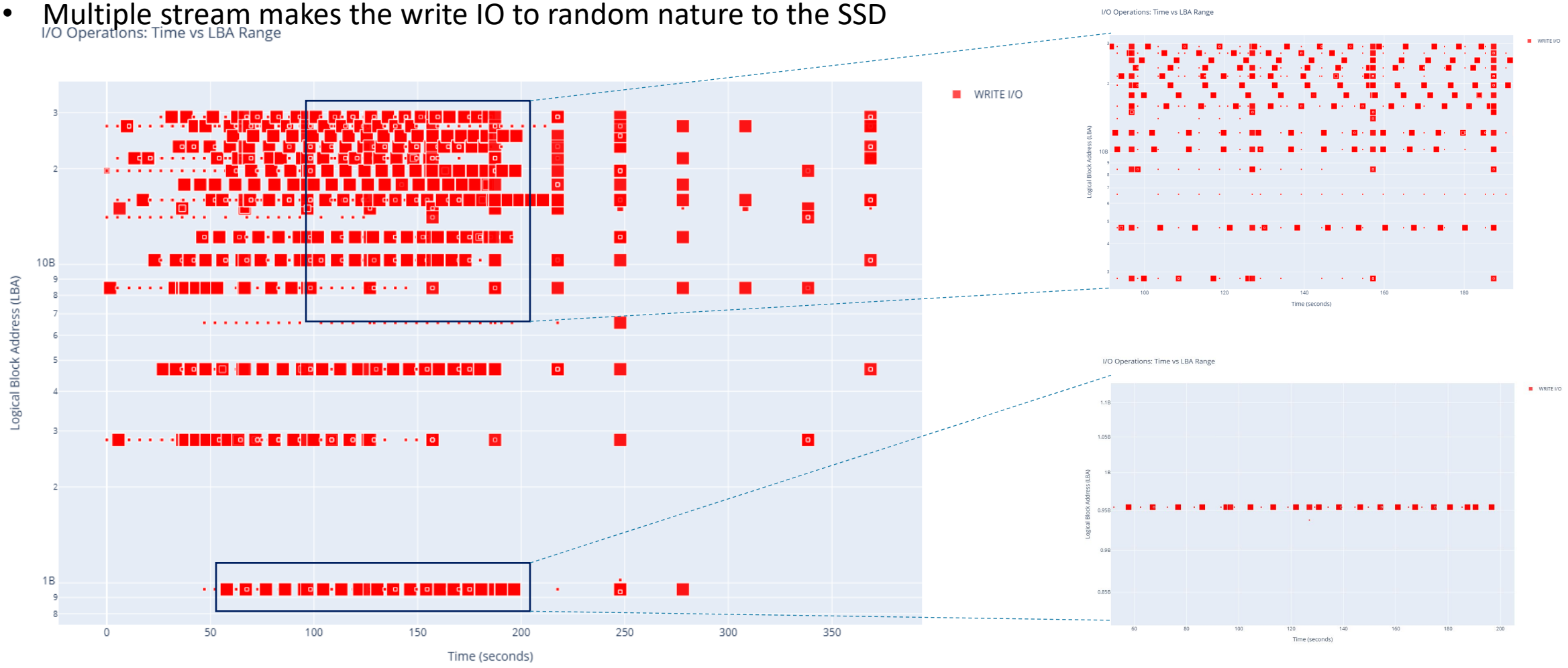
- 128KB (Read): 65%
- 256KB (Read): 33%
- 4 K (write): 0.05%



Analysis for Stable Diffusion: Multi-container Image Generation

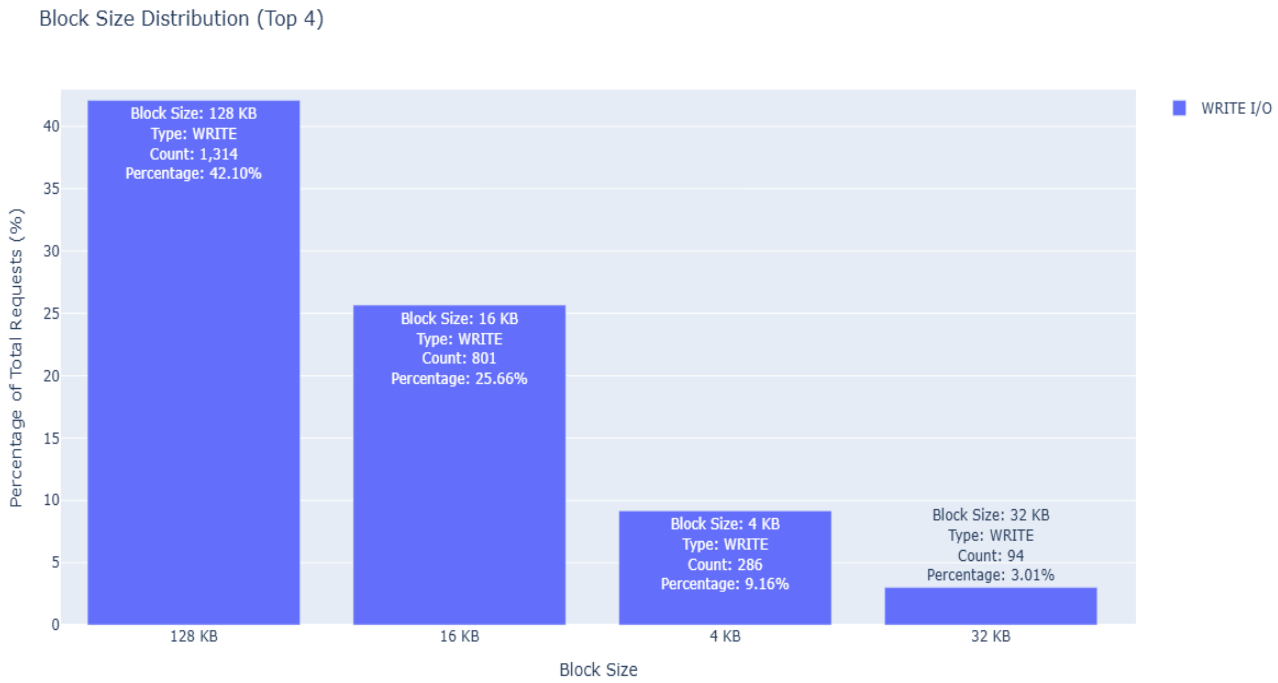
Multiple container:

- Every single container creates image files sequentially
- Multiple stream makes the write IO to random nature to the SSD

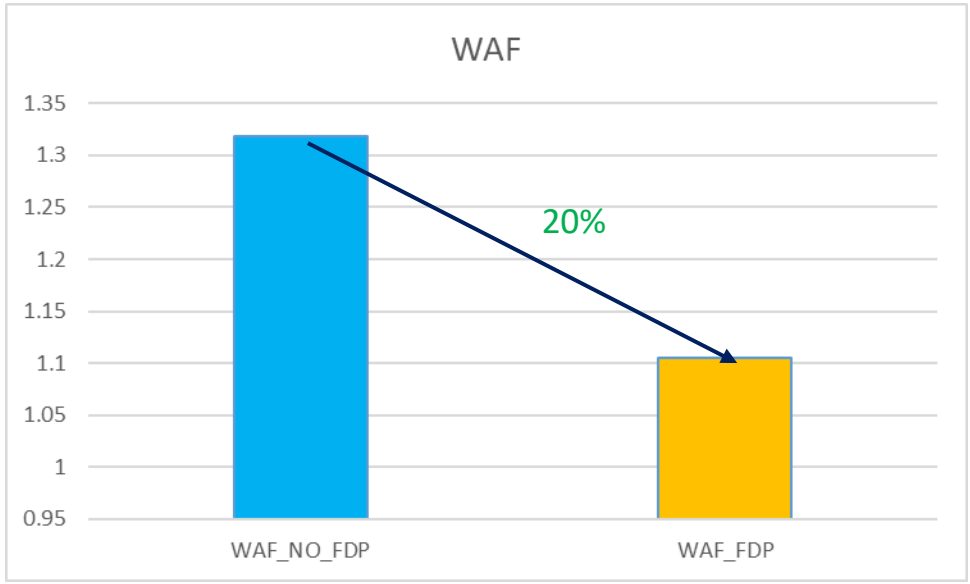


Stable Diffusion on FDP

- **Image generation workload:**
  - **128 K, 16K, 4K, 32K - Access granularity**



Block Distribution - Image Generation



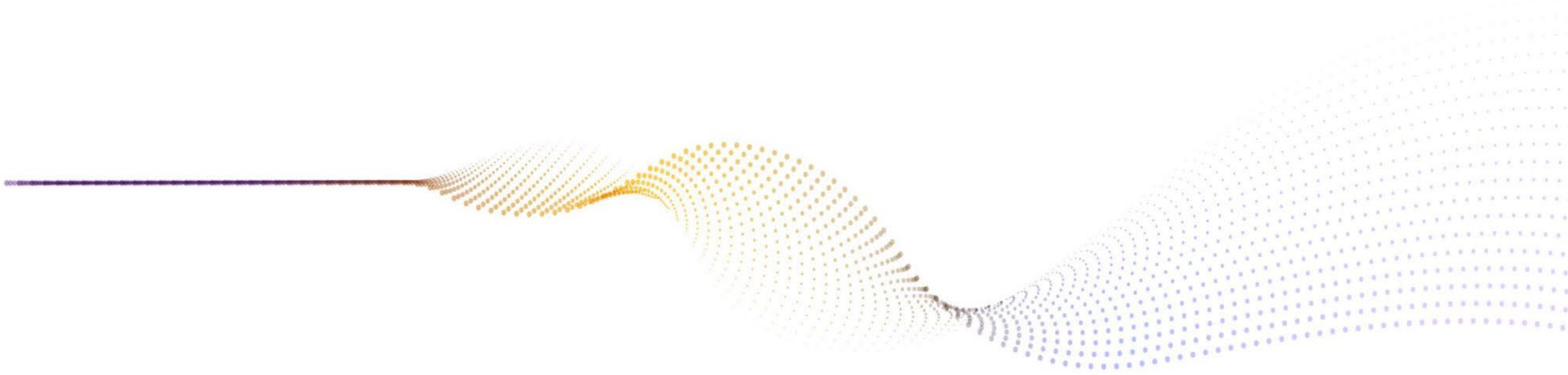
- **SD Workload on FDP enabled and disabled device**
  - **20% WAF reduction**

Summary

- On-Prem Deployment for better control and privacy
- Storage is as critical for better user experience
  - NVMe-oF enables disaggregated, scalable storage
- Random IO is inevitable in AI systems
  - Time to think 100 million IOPS storage
- FDP enabled storage will reduce WAF
- Multi-tier storage setup for performance and capacity
  - High Performance: Latest models and generated images
  - Low Performance: Old models and archived images

Acknowledgements

- SMRC (Samsung Memory Research Center) team for sharing the setup
  - Johnny Kim
  - Seokhyun Ryu
  - Yvette Lee



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

