dstack

От K8S к dstack

Упрощенная оркестрация AI контейнеров в облаке и on-prem

Революция в AI инфраструктуре

- 🚀 **Растущий спрос на инфраструктуру**
- Neoclouds/дата-центры/акселераторы
- OSS инструменты тренировки/инференса
- 🐳 **Контейнеризация**
- **Недостатки K8S/Slurm**

Контейнеры — будущее AI-инфраструктуры

- **Управление зависимостями**
- Ре-использование кода
- 🦖 **Масштабируемость и эффективность**
- 🔒 **Портируемость и приватность**

dstack

Недостатки K8S для AI инфраструктуры

- 🤯 Фокус на DevOps
- **Нет ключевых инструментов для AI**
- 🐌 **Избыточная нагрузка на MLOps**
- Мульти-облака и гибридные среды
- Условная поддержка ускорителей

dstack

Основные абстракции dstack

- Dev environments
- Tasks
- Services
- Fleets

Dev environments

examples/.dstack.yml

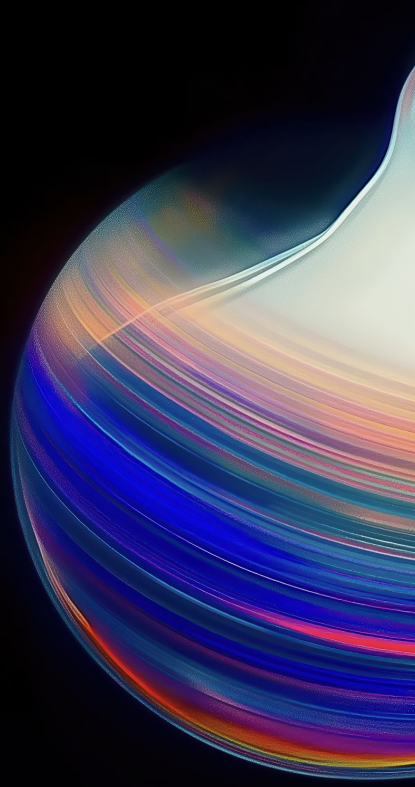
```yaml
type: dev-environment
# The name is optional, if not specified, generated randomly
name: vscode

python: "3.11"
# Uncomment to use a custom Docker image
#image: dstackai/base:py3.13-0.6-cuda-12.1

ide: vscode

# Use either spot or on-demand instances
spot_policy: auto

resources:
  # Required resources
  gpu: 24GB
```

dstack

Tasks

examples/fine-tuning/train.dstack.yml

```yaml
type: task
name: train-distrib

# The size of the cluster
nodes: 2

python: "3.10"
# Commands of the task
commands:
  - pip install -r requirements.txt
  - torchrun --nproc_per_node=$DSTACK_GPUS_PER_NODE
    --node_rank=$DSTACK_NODE_RANK
    --nnodes=$DSTACK_NODES_NUM
    --master_addr=$DSTACK_MASTER_NODE_IP
    --master_port=8008 resnet_ddp.py --num_epochs 20

resources:
  gpu: 24GB
```

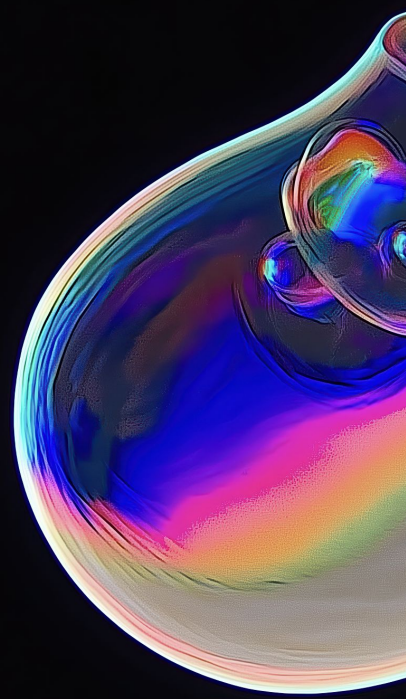


dstack

Services

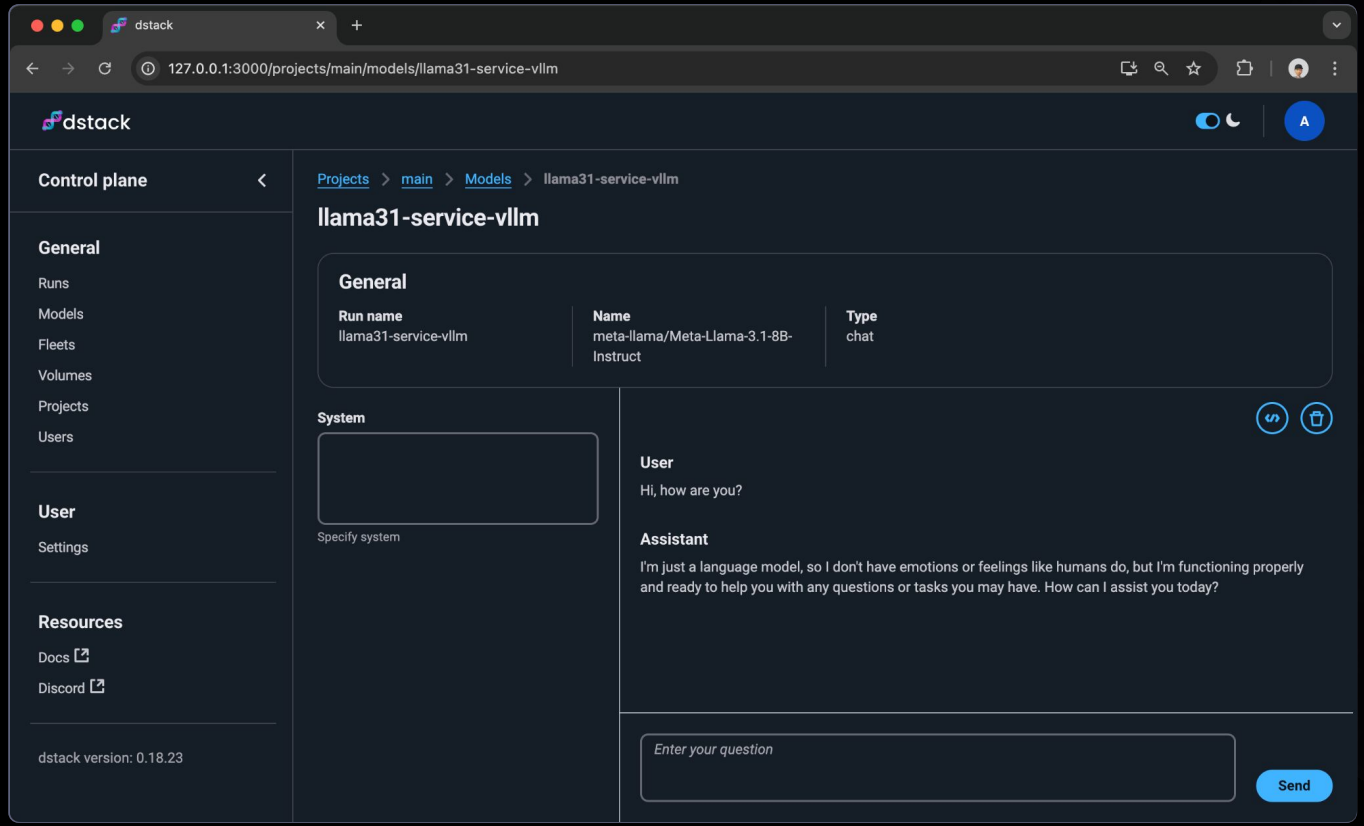
examples/deployment/vllm/service.dstack.yml

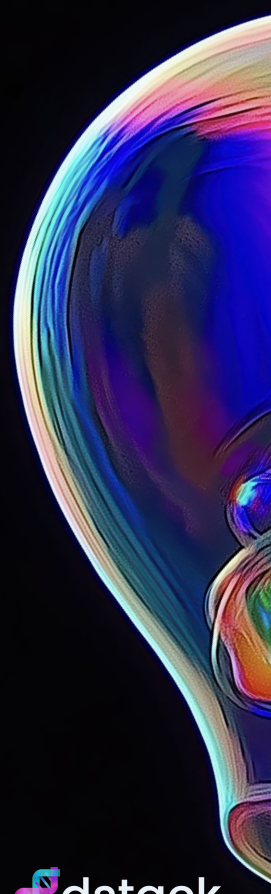
```yaml
type: service
name: llama31-service

python: "3.10"
env:
  - HF_TOKEN
# Commands of the service
commands:
  - pip install vllm
  - vllm serve meta-llama/Meta-Llama-3.1-8B-Instruct --max-model-len 4096
port: 8000

# Register the model
model: meta-llama/Meta-Llama-3.1-8B-Instruct

resources:
  gpu: 24GB
```



dstack

Services





dstack

Fleets (cloud)

examples/misc/fleets/distrib.dstack.yml

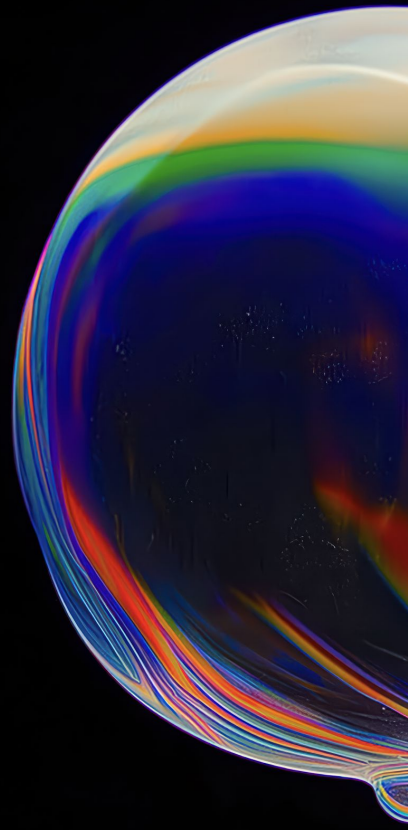
```yaml
type: fleet
# The name is optional, if not specified, generated randomly
name: fleet-distrib

# Number of instances
nodes: 2
# Ensure instances are inter-connected
placement: cluster

# Terminate if idle for 3 days
termination_idle_time: 3d

resources:
  gpu:
    # 24GB or more vRAM
    memory: 24GB..
    # Two or more GPUs
    count: 2..
```

dstack

Fleets (SSH)

examples/misc/fleets/distrib-ssh.dstack.yml

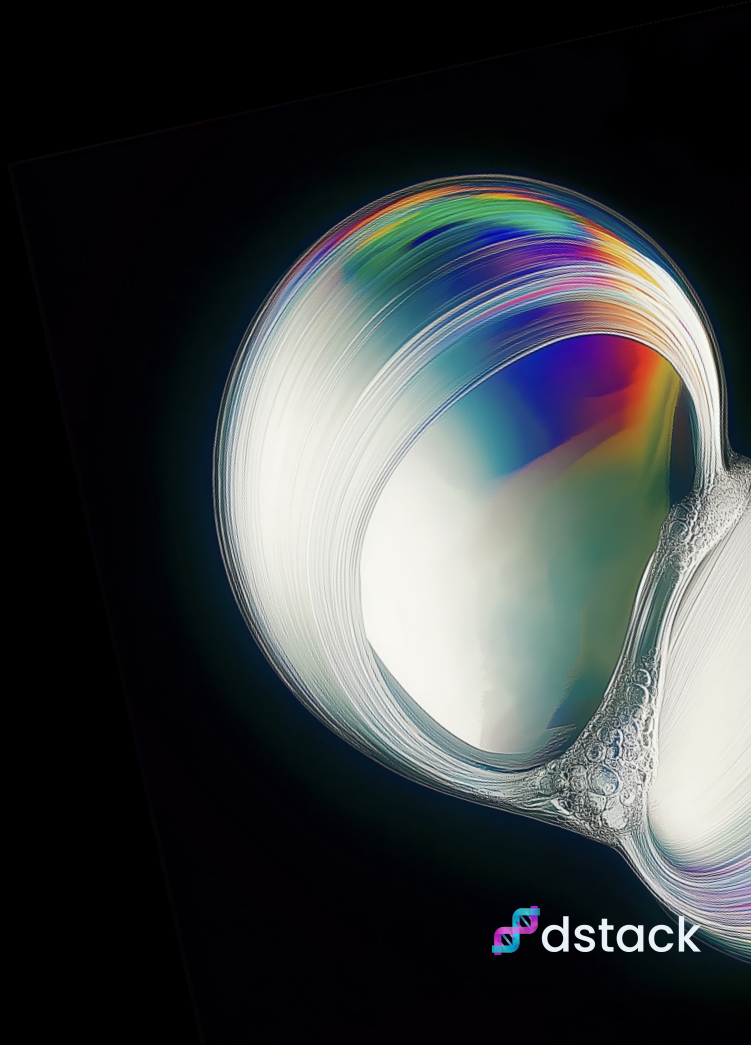
```yaml
type: fleet
# The name is optional, if not specified, generated randomly
name: fleet-distrib-ssh

# Ensure instances are inter-connected
placement: cluster

# The user, private SSH key, and hostnames of the on-prem servers
ssh_config:
  user: ubuntu
  identity_file: ~/.ssh/id_rsa
  hosts:
    - 3.255.177.51
    - 3.255.177.52
```



dstack

Другие абстракции dstack

- Backends
- Gateways
- Volumes
- Projects
- Users



Попробуйте dstack сами

```
$ pip install "dstack[all]" -U
$ dstack server

Applying ~/.dstack/server/config.yml...

The admin token is "bbae0f28-d3dd-4820-bf61-8f4bb40815da"
The server is running at http://127.0.0.1:3000/
```

github.com/dstackai/dstack

dstack