51CTO WOT

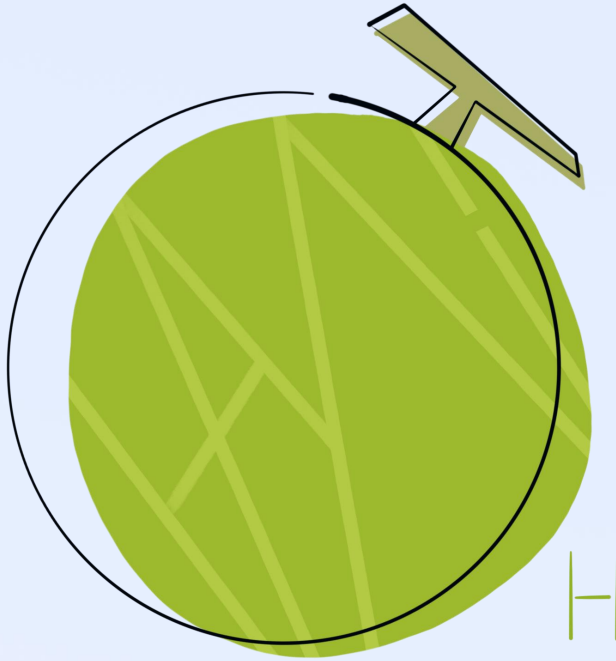
World Of Tech 2024

WOT全球技术创新大会

2024

智启新纪
慧创万物

51CTO WOT



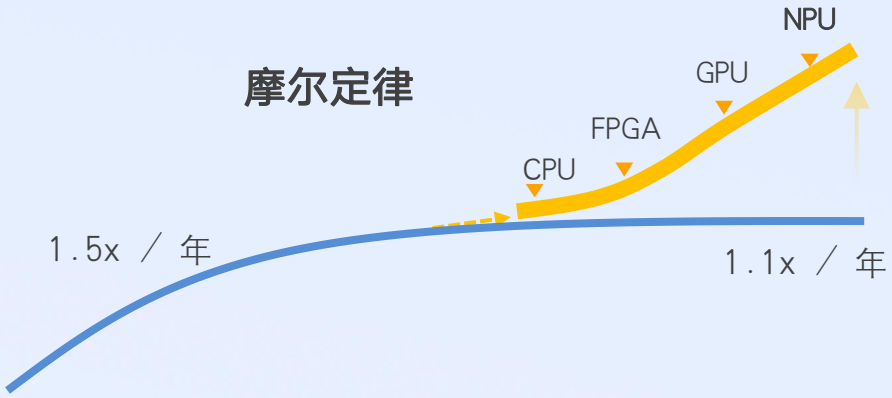
使用虚拟化技术
提升大模型推理性能实践

HAMi

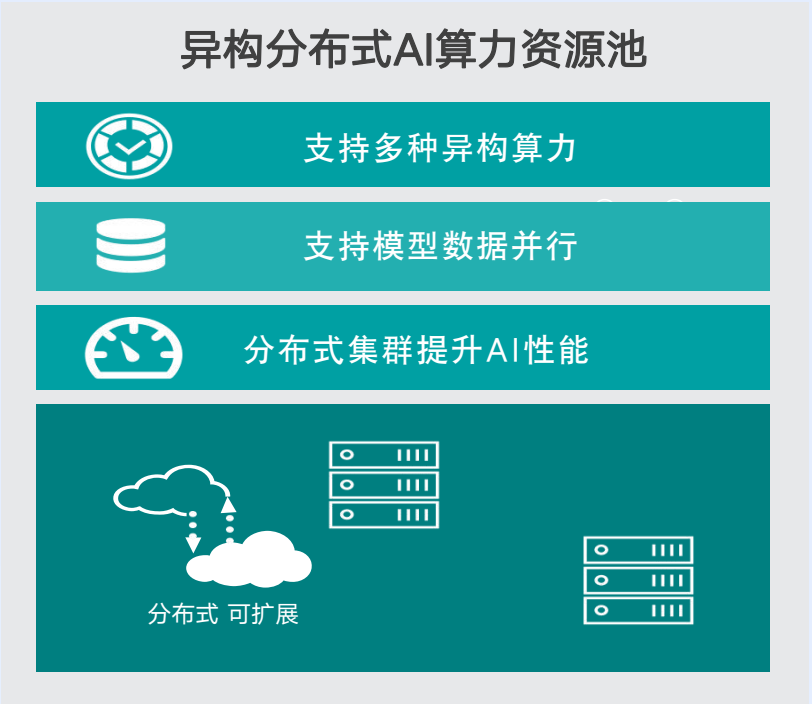
4Paradigm 第四范式 智慧开创 慧创万物 2024

背景1 – 异构分布式AI算力资源池成为必然选择

- 专用芯片/加速卡，能够带来更高效率
- 供应链安全，需要多来源/供应商采购策略
- 信创政策，应对复杂的国家形势变化



单机不再满足
AI算力需求



- 爆发式增长的数据、更大的模型规模、更快的模型更新速度，都对算力带来新的挑战
- CPU的性能从每年提升超过1.5倍降到1.1倍，摩尔定律逐渐失效
- 异构计算架构的创新将打破现有通用计算的瓶颈，推动摩尔定律持续演进

提高训练资源利用率

- GPU池化
- 动态调度感知异构资源
- 自动调整调度策略

提高AI推理资源利用率

- GPU自动划分显存
- 显存超售

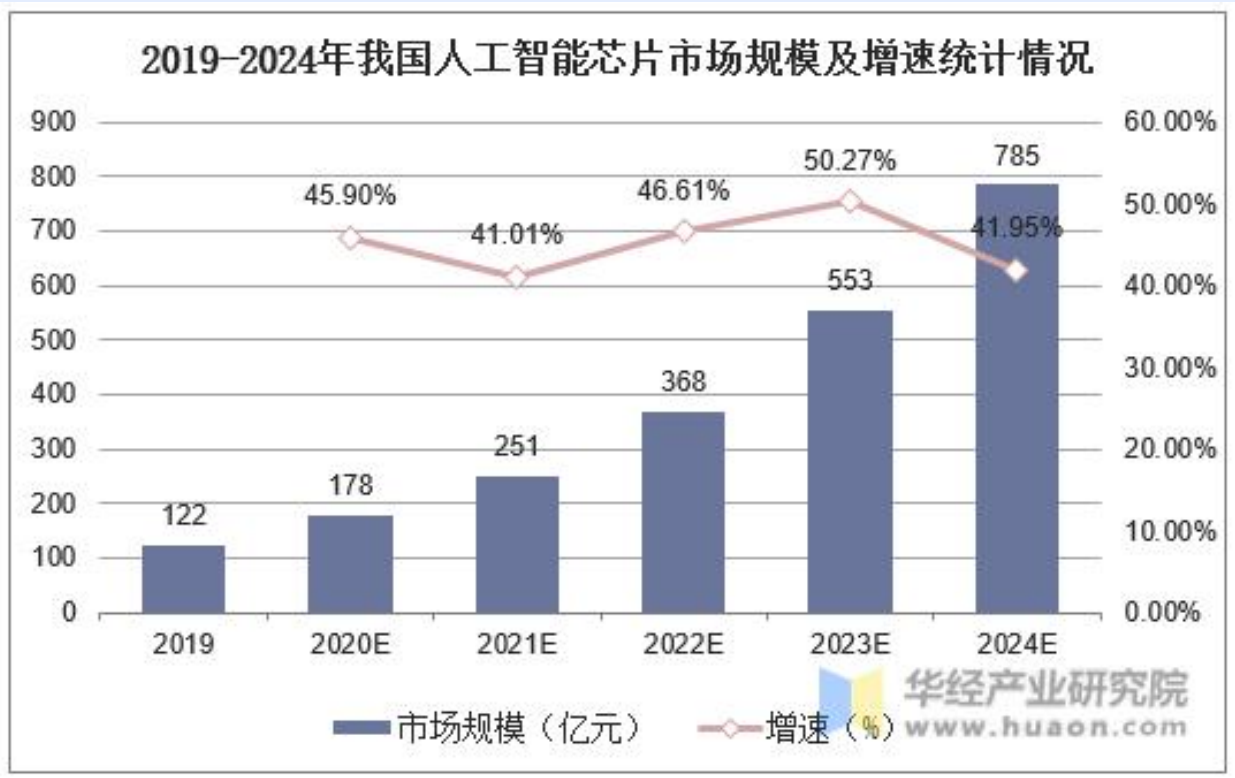
提高任务成功率

- 智能资源配置
- 任务自动配置显存/内存
- 资源精准控制

背景2 – 国产异构算力发展迅猛

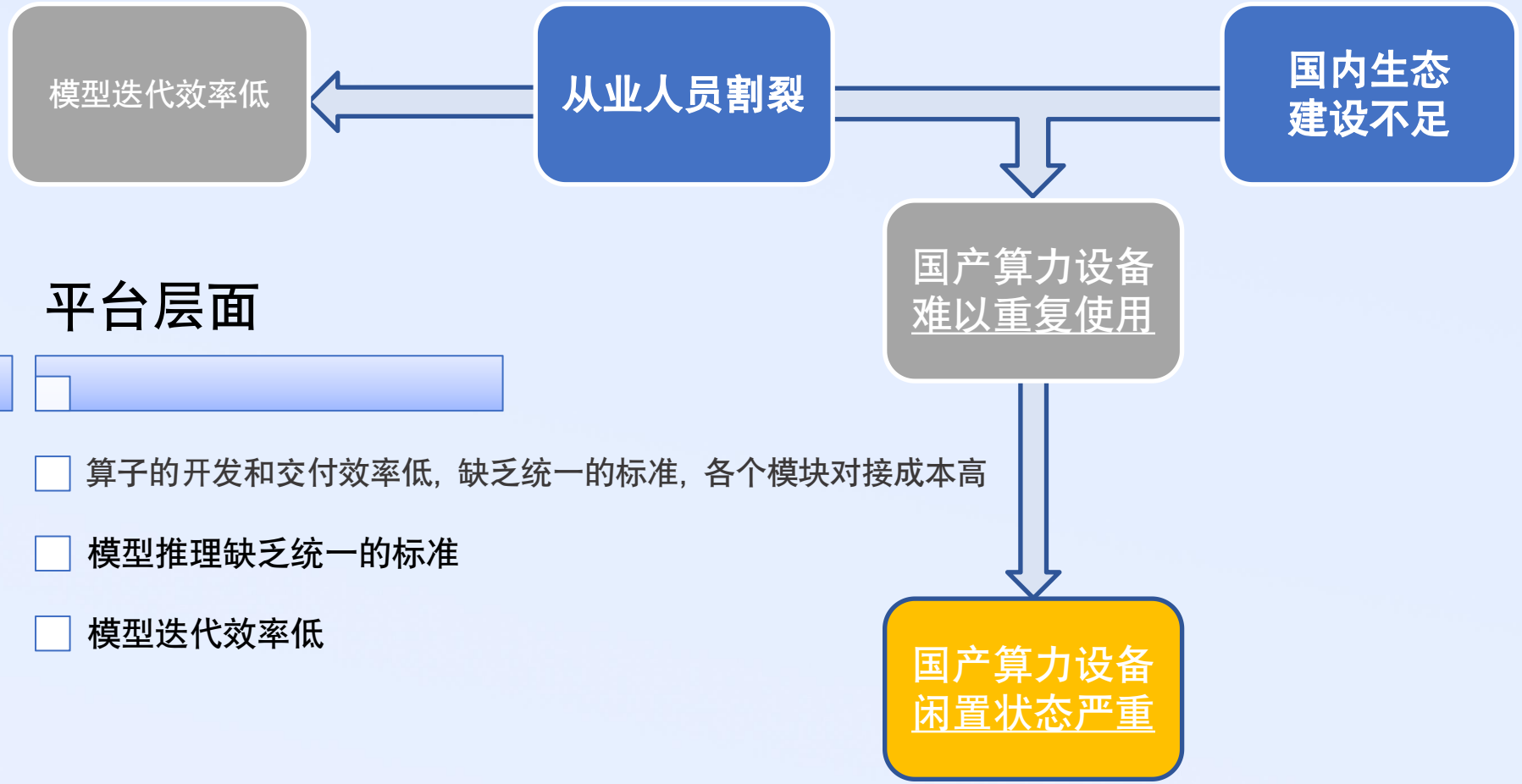
近几年国内不少企业在算力设备方面取得进展，包括海光信息、壁仞科技、燧原科技、摩尔线程等。

- 海光DCU 8000系列，典型功耗260-350W，支持INT4、INT8、FP16、FP32、FP64运算精度，支持4个HBM2内存通道，最高内存带宽为1TB/s、最大内存容量为32GB。海光DCU协处理器全面兼容ROCm GPU计算生态，由于ROCm和CUDA在生态、编程环境等方面具有高度的相似性，CUDA用户可以以较低代价快速迁移至ROCm平台。
- 可以看到，海光DCU是国内唯一支持FP64双精度浮点运算的产品，英伟达的A100、H100都支持FP64，从这一点来看，海光DCU在这方面是比较领先的。
- 天数智芯的BI芯片，集成240亿晶体管，采用7纳米先进制程，支持FP32、FP16、BF16、INT8等多精度数据混合训练，单芯算力每秒147T@FP16。
- 寒武纪2021年11月发布的第三代云端AI芯片思元370，相比于上一代芯片，思元370全面加强了FP16、BF16以及FP32的浮点算力，在全新MLUarch03架构和7nm先进工艺加持下，8位定点算力最高为256TOPS。



Copyright ©2023 4Paradigm All Rights Reserved.

行业痛点



设备层面

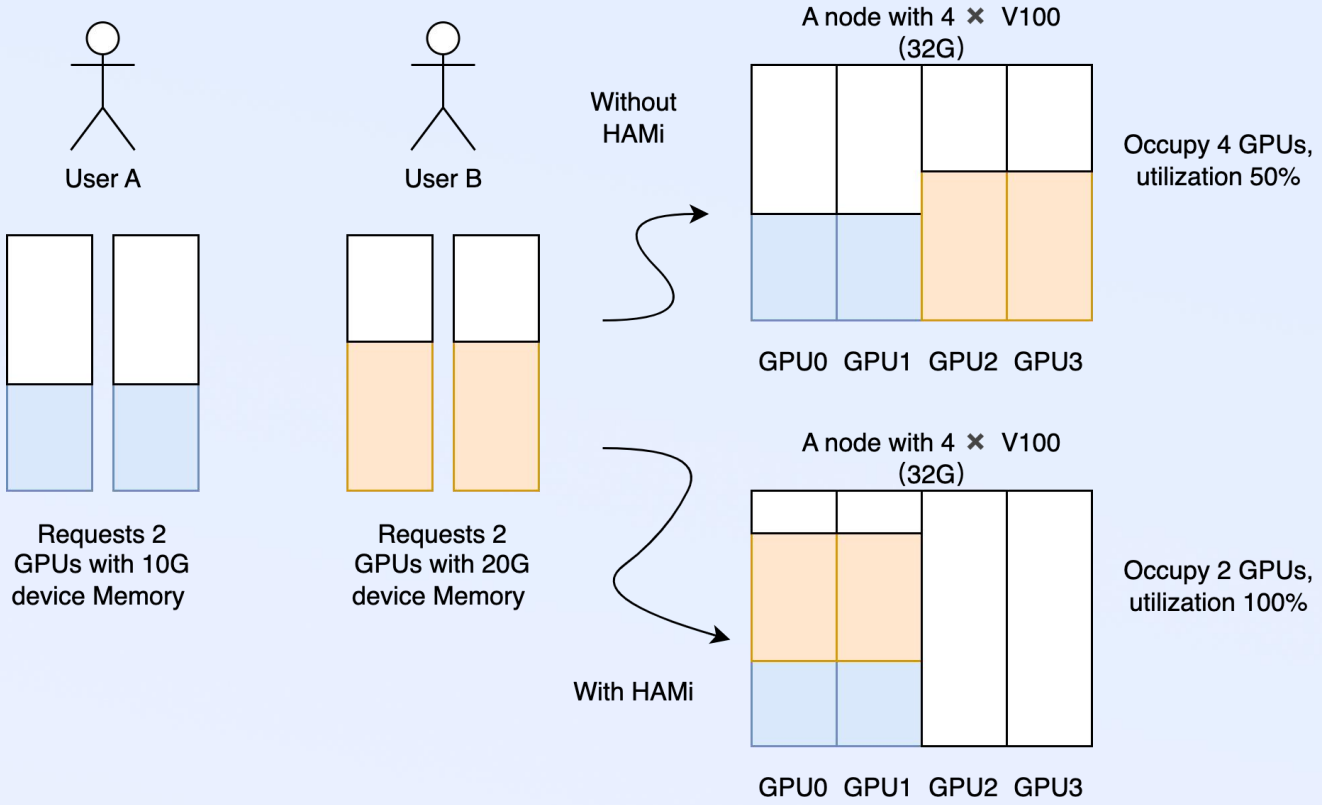
- 国产的生态环境相对封闭
- 从业人员存在严重的割裂
- 国产算力难以重复利用
- 国产算力闲置状况严重

平台层面

- 算子的开发和交付效率低，缺乏统一的标准，各个模块对接成本高
- 模型推理缺乏统一的标准
- 模型迭代效率低

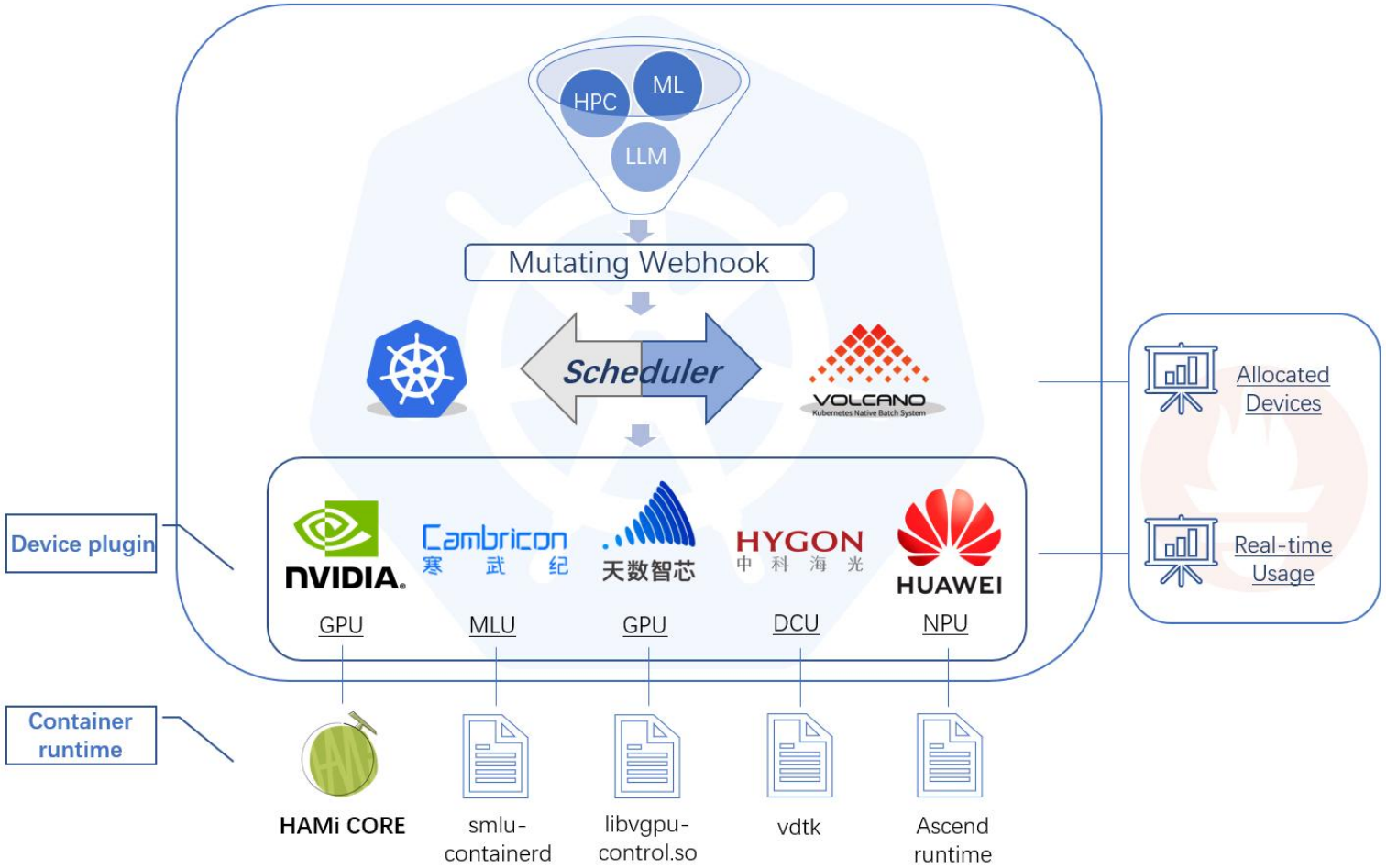
Project-HAMi：基于k8s的算力复用平台

易购算力虚拟化中间件（Heterogeneous AI Computing Virtualization Middleware，简称HAMi，中文名哈密瓜),是一个基于云原声的开源一站式解决不同易购算力复用功能的k8s中间件



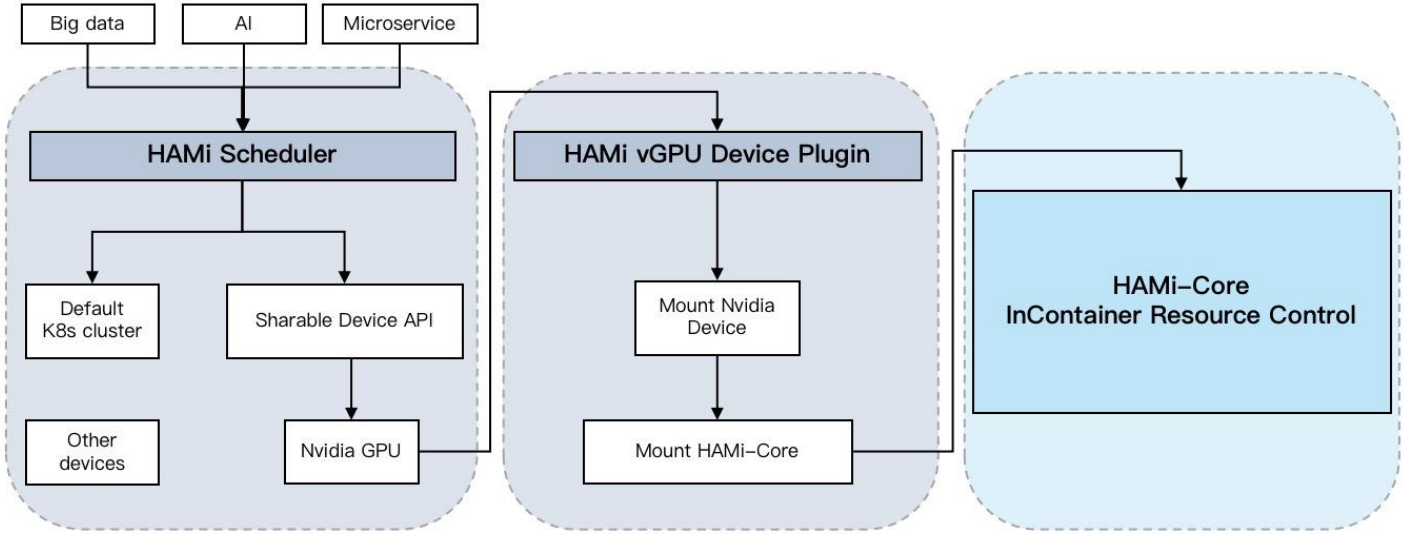
- 大模型经常需要配备一些embedding或者validating功能的小模型，若只能整卡部署，则会造成极大的资源浪费
- 通过虚拟化技术将小模型和大模型复用在一张GPU，从而提升TCO 指标

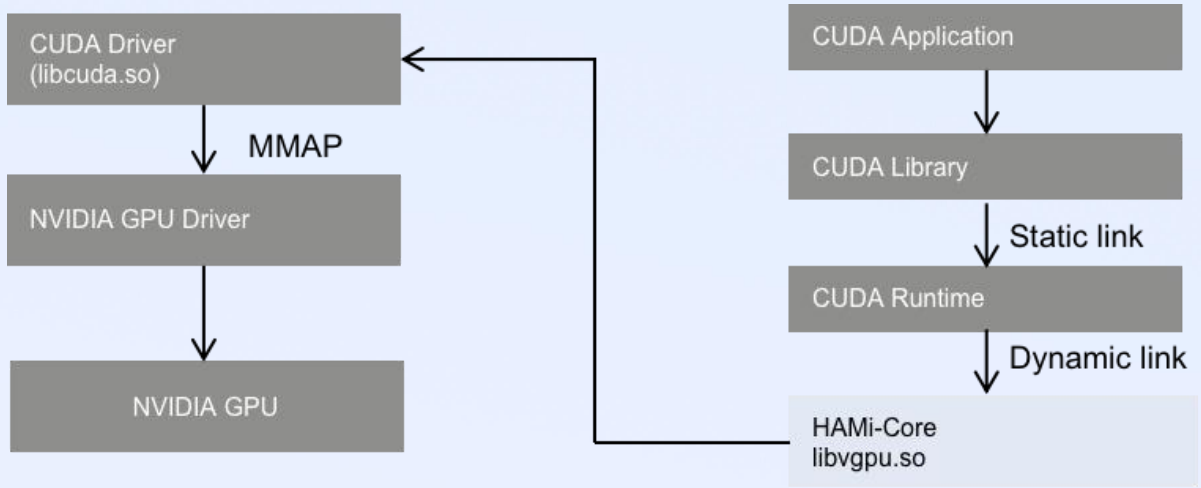
Project-HAMi：架构图



Project-HAMi：架构图

HAMi





HAMi-Core uses symbolic hijacking to operate inside containers

Prerequisites:

- Nvidia driver version >= 440
- CUDA version >= 10.2

Features:

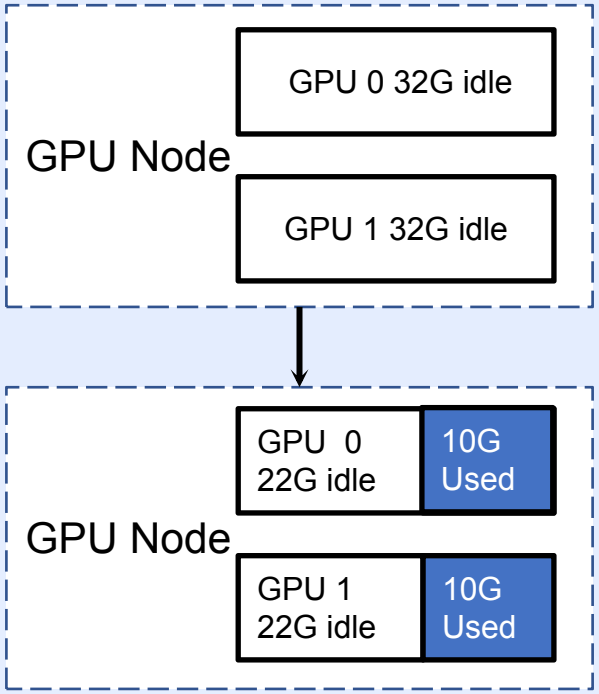
- Device Memory isolation
- Core utilization limitation
- Fault isolation
- Transparent to GPU tasks

使用案例——英伟达

参数描述:

- nvidia.com/gpu: 指定容器中可见的GPU个数.
- nvidia.com/gpumem: 指定每个GPU的显存上限
- nvidia.com/gpucores: 指定每个GPU使用的算力比例

```
$ cat <<EOF | kubectl apply -f -
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod12
spec:
  containers:
    - name: ubuntu-container
      image: ubuntu:18.04
      command: ["bash", "-c", "sleep 86400"]
      resources:
        limits:
              nvidia.com/gpu：2 # requesting 1 vGPUs
              nvidia.com/gpumem：10240
              nvidia.com/gpucores：30
```




```
root@gpu-demo-6b6b88b75b-x9tjb:~# nvidia-smi
[HAMI-core Msg(35:140111864956736:libvgpu.c:836)]: Initializing.....
Thu Apr 18 06:22:54 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.12             Driver Version: 535.104.12   CUDA Version: 12.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA A800 80GB PCIe          On  | 00000000:13:00.0 Off |                    0 |
| N/A   36C    P0              81W / 300W |      0MiB / 10240MiB |      0%      Default |
|                                         |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  NVIDIA A800 80GB PCIe          On  | 00000000:1C:00.0 Off |                    0 |
| N/A   39C    P0              82W / 300W |      0MiB / 10240MiB |      0%      Default |
|                                         |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
[HAMI-core Msg(35:140111864956736:multiprocess_memory_limit.c:434)]: Calling exit handler 35
root@gpu-demo-6b6b88b75b-x9tjb:~#
```

使用案例——寒武纪

- cambricon.com/vmlu: Specifies the number of visible MLUs in the container.
- cambricon.com/mlu.smlu.vmemory: Specifies the memory size to use for each MLU. If not set, the default is to use all available MLU memory.
- nvidia.com/gpucores: Specify the percentage used for each MLU.

```
$ cat <<EOF | kubectl apply -f -

spec:
  containers:
    - ...
      resources:
        limits:
          cambricon.com/vmlu: 1 # requesting 1 vGPUs
          cambricon.com/mlu.smlu.vmemory: 20 # request
20% device memory
          cambricon.com/mlu.smlu.vcore: 10 # request 10%
of compute cores
```

host

```
+-----------------------------------------------------------------------------+
| CNMON v5.10.29                                              Driver v5.10.29 |
+-------------------------------+----------------------+----------------------+
| Card  VF  Name       Firmware |              Bus-Id | Util        Ecc-Error |
| Fan   Temp     Pwr:Usage/Cap |       Memory-Usage | Mode     Compute-Mode |
|===============================+======================+======================|
| 0     /   MLU370-X4    v1.1.6 |      0000:B4:00.0 | 0%                 N/A |
|  0%   35C        28 W/ 150 W |    0 MiB/ 23308 MiB | FULL           Default |
+-------------------------------+----------------------+----------------------+
```

container

```
root@binpack-1-5bff578c96-bbnwk:/# cnmon
Fri Apr 19 10:20:44 2024
+-----------------------------------------------------------------------------+
| CNMON v5.10.29                                              Driver v5.10.29 |
+-------------------------------+----------------------+----------------------+
| Card  VF  Name       Firmware |              Bus-Id | Util        Ecc-Error |
| Fan   Temp     Pwr:Usage/Cap |       Memory-Usage | Mode     Compute-Mode |
|===============================+======================+======================|
| 0     /   MLU370-X4    v1.1.6 |      0000:B5:00.0 | N/A                N/A |
|  0%   34C        27 W/ 150 W |               N/A | sMLU           Default |
+-------------------------------+----------------------+----------------------+


+-----------------------------------------------------------------------------+
| sMLU Info:                                                                  |
+-------------------------+-----------------------+---------------------------+
| Card   Profile  Instance |                BDF | Usage                     |
|      ID       ID        |       Memory-Usage |                       MLU |
+-------------------------+-----------------------+---------------------------+
| 0      0        1       |       0000:B5:00.0 |                        0% |
|                         |    0 MiB/ 4660 MiB |                           |
+-------------------------+-----------------------+---------------------------+


+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  Card  MI  PID     Command Line                         MLU Memory Usage |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

使用案例——天数智芯

- iluvatar.ai/gpu: Specifies the number of visible iluvatar GPUs in the container.
- iluvatar.ai/vcuda-memory: Specifies the memory size to use for each iluvatar GPU. If not set, the default is to use all available device memory.
- iluvatar.ai/vcuda-core: Specify the percentage used for each Iluvatar GPU.

host

```
+-----------------------------------------------------------------------------+
| IX-ML: 4.0.0       Driver Version: 4.0.0       CUDA Version: 10.2           |
|-------------------------------+----------------------+----------------------|
| GPU  Name                     | Bus-Id               | Clock-SM  Clock-Mem  |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage    | GPU-Util  Compute M. |
|===============================+======================+======================|
| 0    Iluvatar MR-V100         | 00000000:40:00.0     | 1500MHz   1600MHz    |
| 0%   44C   P0    41W / 150W   | 114MiB / 32768MiB    | 0%        Default    |
+-------------------------------+----------------------+----------------------+
```

container

```
$ cat <<EOF | kubectl apply -f -

spec:
  containers:
    - ...
      resources:
        limits:
          iluvatar.ai/gpu: 1
          iluvatar.ai/vcuda-core: 50
          iluvatar.ai/vcuda-memory: 64 #each unit
represents 256M device memory
```

```
[root@poddemo corex-3.2.0]# ixsmi
Timestamp    Tue Feb 20 10:38:32 2024
+-----------------------------------------------------------------------------+
| IX-ML: 3.2.0.2     Driver Version: 3.2.0       CUDA Version: 10.2           |
|-------------------------------+----------------------+----------------------|
| GPU  Name                     | Bus-Id               | Clock-SM  Clock-Mem  |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage    | GPU-Util  Compute M. |
|===============================+======================+======================|
| 0    Iluvatar MR-V100         | 00000000:40:00.0     | 1500MHz   1600MHz    |
| 0%   43C   P0    40W / 150W   | 0MiB / 16384MiB      | 0%        Default    |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID      Process name                                 Usage(MiB) |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

使用案例——华为升腾910B

- huawei.com/ascend910: Specifies the number of visible Ascend 910s in the container.
- huawei.com/ascend910-memory: Specifies the memory size to use for each Ascend 910s. If not set, the default is to use all available device memory.

host

```
+-------------------------------------------------------------------------------------------+
| npu-smi 24.1.rc1                 Version: 24.1.rc1                                        |
+---------------------------+---------------+---------------------------------------------+
| NPU   Name                | Health        | Power(W)    Temp(C)           Hugepages-Usage(page)|
| Chip                      | Bus-Id        | AICore(%)   Memory-Usage(MB)  HBM-Usage(MB)        |
+===========================+===============+=============================================+
| 0     910B3               | OK            | 95.9        39                0    / 0          |
| 0                         | 0000:C1:00.0  | 0           0    / 0          24082/ 65536         |
+---------------------------+---------------+---------------------------------------------+
```

container

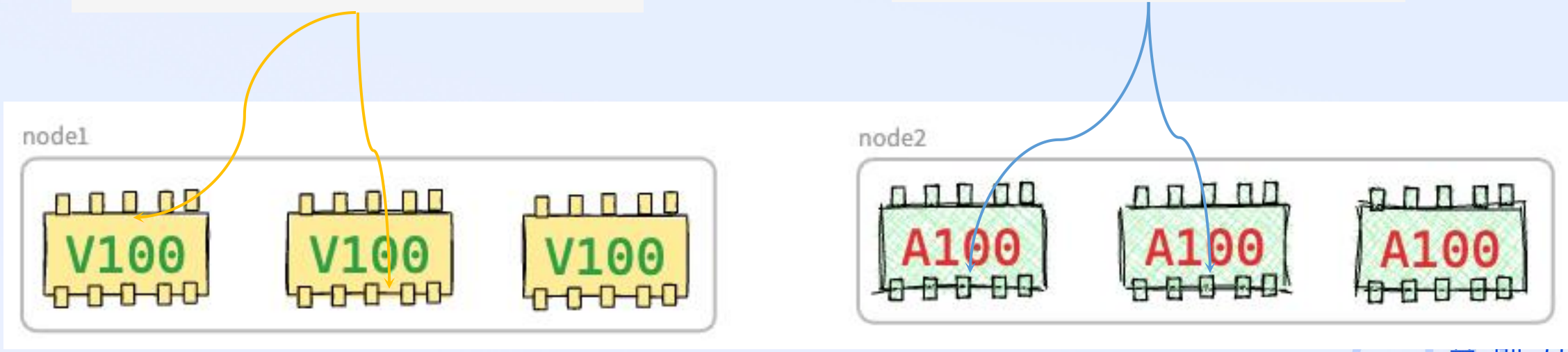
```
$ cat <<EOF | kubectl apply -f -

spec:
  containers:
    - ...
      resources:
        limits:
          huawei.com/Ascend910: 1
          huawei.com/Ascend910-memory: 16384
```

```
+-------------------------------------------------------------------------------------------+
| npu-smi 24.1.rc1                 Version: 24.1.rc1                                        |
+---------------------------+---------------+---------------------------------------------+
| NPU   Name                | Health        | Power(W)    Temp(C)           Hugepages-Usage(page)|
| Chip                      | Bus-Id        | AICore(%)   Memory-Usage(MB)  HBM-Usage(MB)        |
+===========================+===============+=============================================+
| 1     910B3vir05_1c_16g   | OK            | 89.9        25                0    / 0          |
| 0                         | 0000:C2:00.4  | 0           0    / 0          1235 / 16384         |
+===========================+===============+=============================================+
+---------------------------+---------------+---------------------------------------------+
| NPU     Chip              | Process id    | Process name                | Process memory(MB)      |
+===========================+===============+=============================================+
| No running processes found in NPU 1                                                       |
+===========================+===============+=============================================+
```

Project-HAMi：指定设备种类

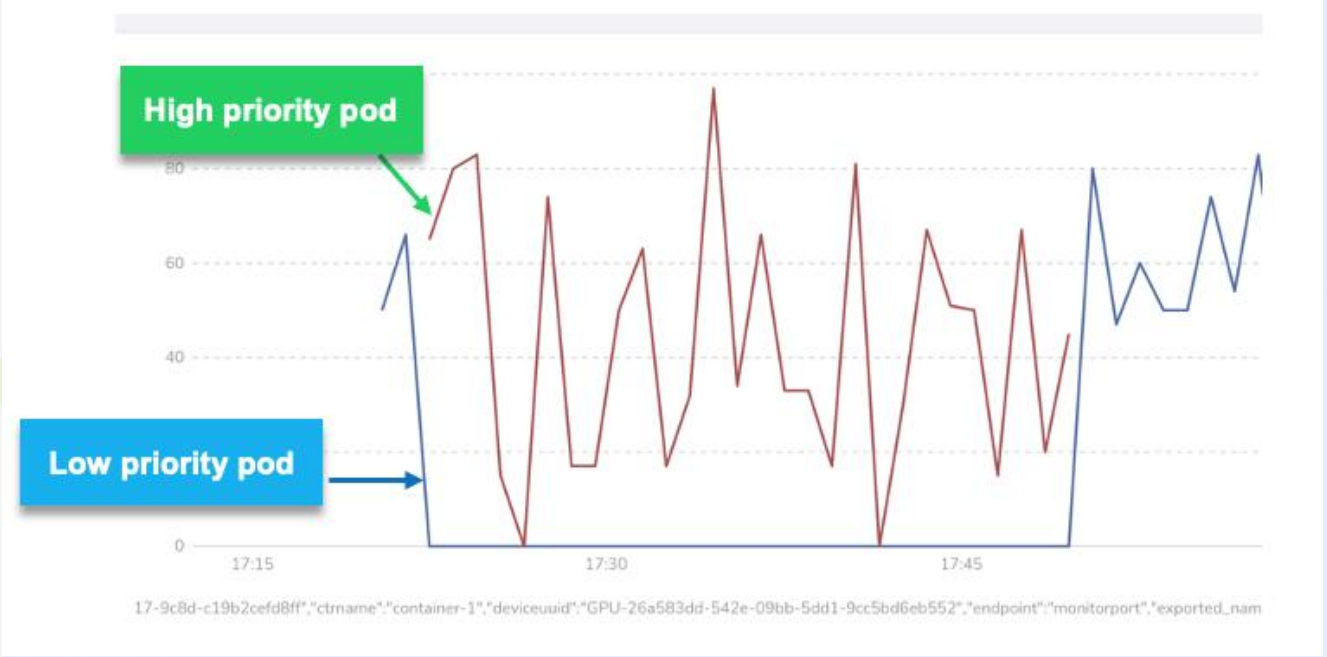
```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod2
  annotations:
    nvidia.com/nouse-gputype: "A100"
spec:
  containers:
    - name: ubuntu-container
      image: ubuntu:18.04
      command: ["bash", "-c", "sleep 86400"]
      resources:
        limits:
          nvidia.com/gpu: 2
```

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
  annotations:
    nvidia.com/use-gputype: "A100"
spec:
  containers:
    - name: ubuntu-container
      image: ubuntu:18.04
      command: ["bash", "-c", "sleep 86400"]
      resources:
        limits:
          nvidia.com/gpu: 2
```



Project-HAMi 算力超售与抢占

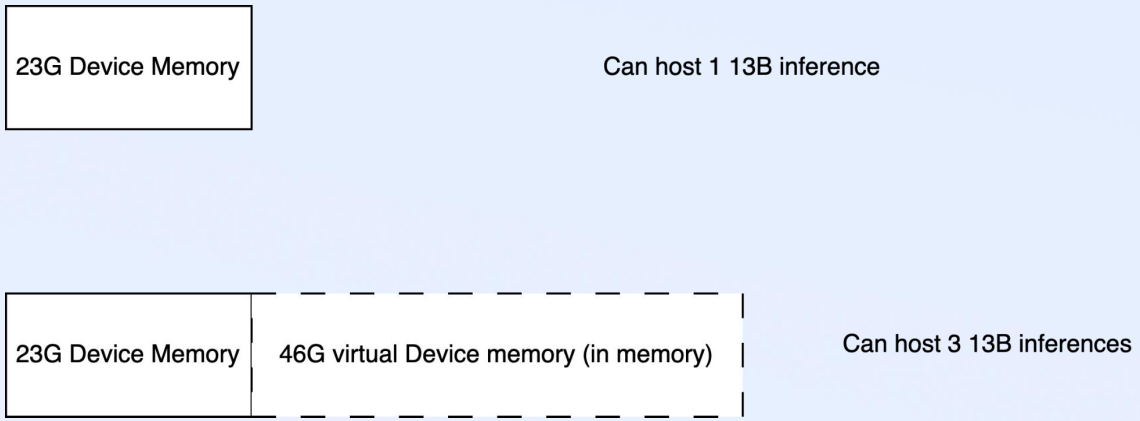
```yaml
apiVersion: v1
kind: Pod
metadata:
  name: gpu-pod
spec:
  containers:
    - name: ubuntu-container
      image: ubuntu:18.04
      command: ["bash", "-c", "sleep 86400"]
      env:
      # vgpu task priority 0 for high and 1 for low
        - name: CUDA_TASK_PRIORITY'
          value: '0'
      resources:
        limits:
          nvidia.com/gpu: 1
          nvidia.com/gpumem: 3000
          nvidia.com/gpucores: 30
```



Project-HAMi 显存超售

显存超售支持:

通过配置 'deviceMemoryScaling>1' 即可激活虚拟显存,例如在部署时指定`deviceMemoryScaling=3`就会把每张卡的显存大小扩大到3倍




```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-PCIE...  Off  | 00000000:1B:00.0 Off |                    0 |
| N/A   52C    P0    39W / 250W |  32501MiB / 32510MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-PCIE...  Off  | 00000000:88:00.0 Off |                    0 |
| N/A   50C    P0    42W / 250W |  28061MiB / 32510MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     12312      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     12679      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     13296      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     14622      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     15495      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     16247      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     16875      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     17828      C   ./bin/nnpredictor                817MiB |
|    0   N/A  N/A     18602      C   ./bin/nnpredictor                817MiB |
|    1   N/A  N/A     65438      C   ./bin/nnpredictor              28057MiB |
+-----------------------------------------------------------------------------+
```

Project-HAMi 算力隔离

可以通过指定nvidia.com/gpucores来达到算力隔离的效果

```
kind: Pod
…
spec:
  containers:
    - …
      resources:
        limits:
          nvidia.com/gpu: 1 # requesting 1 vGPUs
                nvidia.com/gpucores：100 # request 100% compute
cores
```

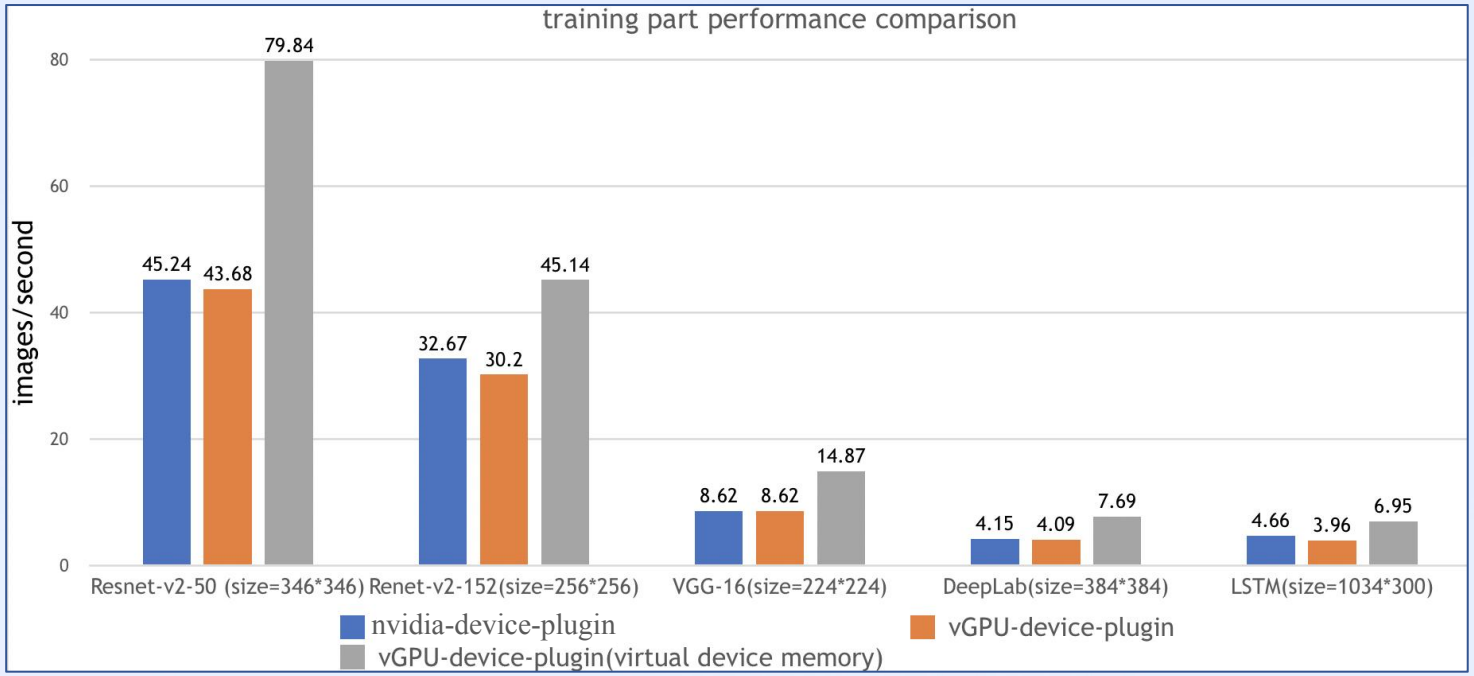
```
images/sec: 320.7 +/- 0.0 (jitter = 0.0)      7.765
images/sec: 319.9 +/- 0.4 (jitter = 0.3)      8.049
images/sec: 318.7 +/- 0.4 (jitter = 2.3)      7.808
images/sec: 318.6 +/- 0.3 (jitter = 2.1)      7.976
images/sec: 318.9 +/- 0.3 (jitter = 1.8)      7.591
images/sec: 319.0 +/- 0.2 (jitter = 1.8)      7.549
images/sec: 318.9 +/- 0.2 (jitter = 1.7)      7.819
images/sec: 318.5 +/- 0.3 (jitter = 1.7)      7.820
images/sec: 318.5 +/- 0.4 (jitter = 1.4)      7.847
images/sec: 318.4 +/- 0.3 (jitter = 1.6)      8.027
```

```
kind: Pod
…
spec:
  containers:
    - …
      resources:
        limits:
          nvidia.com/gpu: 1 # requesting 1 vGPUs
                nvidia.com/gpucores：60 # request 60% compute
cores
```

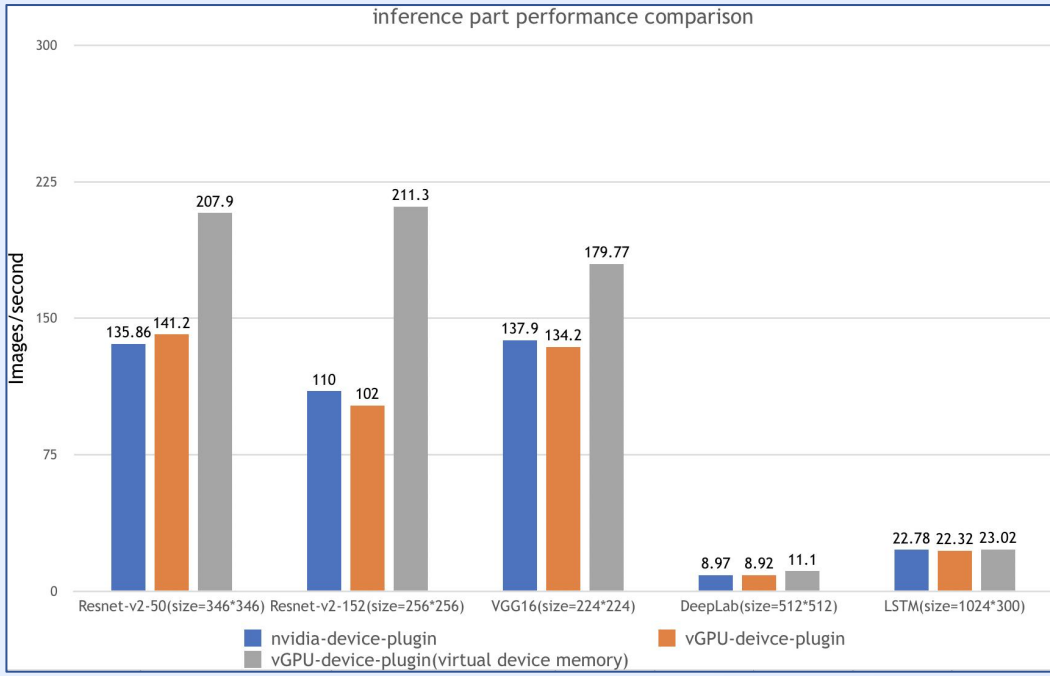
```
images/sec: 188.9 +/- 1.6 (jitter = 6.6)      7.737
images/sec: 188.8 +/- 1.6 (jitter = 6.6)      7.616
images/sec: 189.0 +/- 1.6 (jitter = 6.6)      7.700
images/sec: 188.9 +/- 1.6 (jitter = 6.6)      7.384
images/sec: 188.7 +/- 1.6 (jitter = 6.6)      7.762
images/sec: 189.0 +/- 1.6 (jitter = 6.6)      7.834
images/sec: 188.9 +/- 1.6 (jitter = 6.6)      7.574
images/sec: 188.8 +/- 1.6 (jitter = 6.6)      7.672
images/sec: 189.0 +/- 1.6 (jitter = 6.6)      7.438
images/sec: 188.8 +/- 1.6 (jitter = 6.6)      7.721
```

Project-HAMi 性能

训练



推理



Test Environment：

- GPU Type：Tesla V100
- GPU Num：1
- Kubernetes Version：v1.12.9
- Docker Version：v18.09.1

Test Instance：

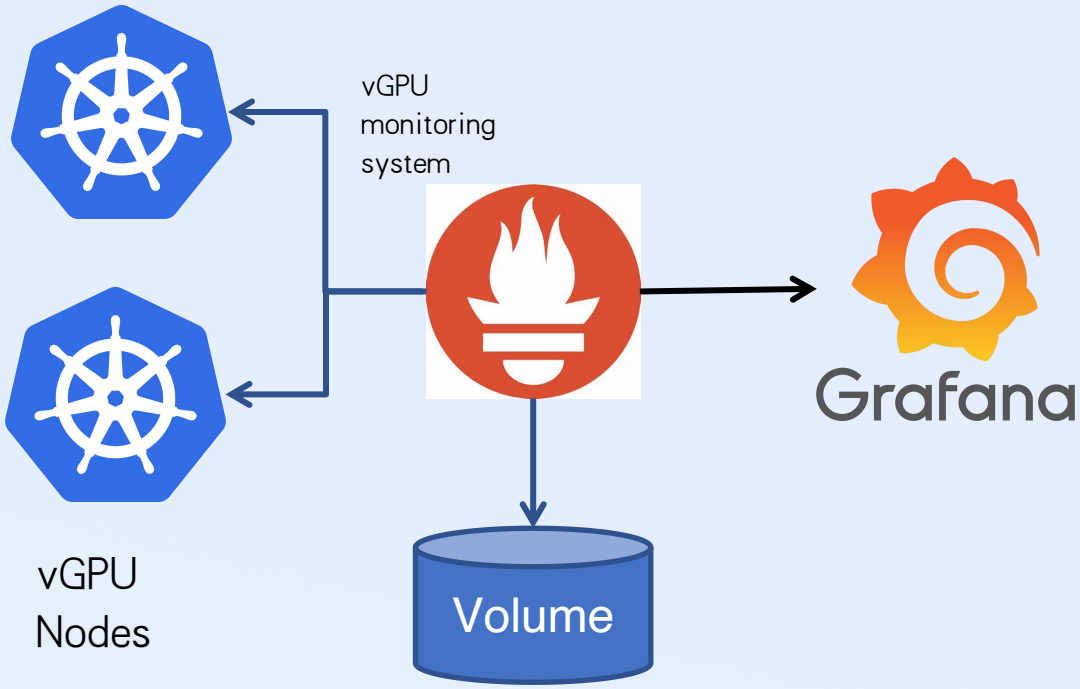
- nvidia-device-plugin：基于Nvidia源生device plugin在1块GPU上运行1个任务/服务
- vGPU-device-plugin：基于第四范式vGPU device plugin在1块vGPU上运行1个任务/服务
- vGPU-device-plugin(virtual device memory)：基于第四范式vGPU device plugin在2块vGPU上运行2个任务/服务

Project-HAMi 监控接口

HAMi

Not Secure | 172.26.1.29:31993/metrics

```
# HELP GPUDeviceCoreAllocated Device core allocated for a certain GPU
# TYPE GPUDeviceCoreAllocated gauge
GPUDeviceCoreAllocated{deviceidx="0",deviceuuid="DCU-0",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="0",deviceuuid="GPU-00552014-5c87-89ac-b1a6-7b53aa24b0ec",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="0",deviceuuid="MLU-45013011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="1",deviceuuid="DCU-1",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="1",deviceuuid="GPU-0fc3eda5-e98b-a25b-5b0d-cf5c855d1448",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="1",deviceuuid="MLU-54043011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="2",deviceuuid="DCU-2",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="2",deviceuuid="MLU-07053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="3",deviceuuid="DCU-3",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="3",deviceuuid="MLU-71053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="4",deviceuuid="DCU-4",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="5",deviceuuid="DCU-5",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="6",deviceuuid="DCU-6",nodeid="node1",zone="vGPU"} 0
GPUDeviceCoreAllocated{deviceidx="7",deviceuuid="DCU-7",nodeid="node1",zone="vGPU"} 60
# HELP GPUDeviceCoreLimit Device memory core limit for a certain GPU
# TYPE GPUDeviceCoreLimit gauge
GPUDeviceCoreLimit{deviceidx="0",deviceuuid="DCU-0",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="0",deviceuuid="GPU-00552014-5c87-89ac-b1a6-7b53aa24b0ec",nodeid="node67-4v100",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="0",deviceuuid="MLU-45013011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreLimit{deviceidx="1",deviceuuid="DCU-1",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="1",deviceuuid="GPU-0fc3eda5-e98b-a25b-5b0d-cf5c855d1448",nodeid="node67-4v100",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="1",deviceuuid="MLU-54043011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreLimit{deviceidx="2",deviceuuid="DCU-2",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="2",deviceuuid="MLU-07053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreLimit{deviceidx="3",deviceuuid="DCU-3",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="3",deviceuuid="MLU-71053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceCoreLimit{deviceidx="4",deviceuuid="DCU-4",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="5",deviceuuid="DCU-5",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="6",deviceuuid="DCU-6",nodeid="node1",zone="vGPU"} 100
GPUDeviceCoreLimit{deviceidx="7",deviceuuid="DCU-7",nodeid="node1",zone="vGPU"} 100
# HELP GPUDeviceMemoryAllocated Device memory allocated for a certain GPU
# TYPE GPUDeviceMemoryAllocated gauge
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="0",deviceuuid="DCU-0",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="0",deviceuuid="GPU-00552014-5c87-89ac-b1a6-7b53aa24b0ec",nodeid="node67-4v100",zone="vGPU"} 3.43
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="0",deviceuuid="MLU-45013011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="1",deviceuuid="DCU-1",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="1",deviceuuid="GPU-0fc3eda5-e98b-a25b-5b0d-cf5c855d1448",nodeid="node67-4v100",zone="vGPU"} 3.43
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="1",deviceuuid="MLU-54043011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="2",deviceuuid="DCU-2",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="2",deviceuuid="MLU-07053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="3",deviceuuid="DCU-3",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="3",deviceuuid="MLU-71053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 2.09
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="4",deviceuuid="DCU-4",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="5",deviceuuid="DCU-5",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="0",deviceidx="6",deviceuuid="DCU-6",nodeid="node1",zone="vGPU"} 0
GPUDeviceMemoryAllocated{devicecores="60",deviceidx="7",deviceuuid="DCU-7",nodeid="node1",zone="vGPU"} 2.097152e+09
# HELP GPUDeviceMemoryLimit Device memory limit for a certain GPU
# TYPE GPUDeviceMemoryLimit gauge
GPUDeviceMemoryLimit{deviceidx="0",deviceuuid="DCU-0",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="0",deviceuuid="GPU-00552014-5c87-89ac-b1a6-7b53aa24b0ec",nodeid="node67-4v100",zone="vGPU"} 3.4359738368e+10
GPUDeviceMemoryLimit{deviceidx="0",deviceuuid="MLU-45013011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 2.4440209408e+10
GPUDeviceMemoryLimit{deviceidx="1",deviceuuid="DCU-1",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="1",deviceuuid="GPU-0fc3eda5-e98b-a25b-5b0d-cf5c855d1448",nodeid="node67-4v100",zone="vGPU"} 3.4359738368e+10
GPUDeviceMemoryLimit{deviceidx="1",deviceuuid="MLU-54043011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 2.4440209408e+10
GPUDeviceMemoryLimit{deviceidx="2",deviceuuid="DCU-2",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="2",deviceuuid="MLU-07053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 2.4440209408e+10
GPUDeviceMemoryLimit{deviceidx="3",deviceuuid="DCU-3",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="3",deviceuuid="MLU-71053011-2257-0000-0000-000000000000",nodeid="node67-4v100",zone="vGPU"} 2.4440209408e+10
GPUDeviceMemoryLimit{deviceidx="4",deviceuuid="DCU-4",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="5",deviceuuid="DCU-5",nodeid="node1",zone="vGPU"} 3.4342961152e+10
GPUDeviceMemoryLimit{deviceidx="6",deviceuuid="DCU-6",nodeid="node1",zone="vGPU"} 3.4342961152e+10
```

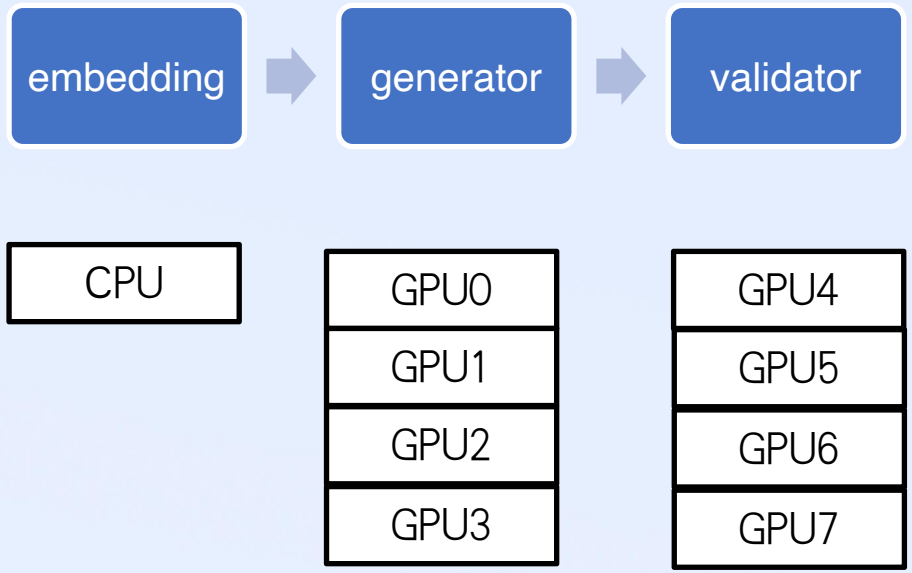


实践案例：第四范式推理加速框架SLX LLM

第四范式发布了大模型推理框架SLXLLM以及推理加速卡SLX，在二者联合优化下，在文本生成类场景中，大模型推理性能提升10倍。例如在使用4张80G GPU对72B大模型进行推理测试中，相较于使用vLLM，第四范式使用SLXLLM+SLX的方案。

- 可同时运行任务数量从4增至40。
- 可兼容TGI、FastLLM、vLLM等主流大模型推理框架
- 大模型推理性能提升约1–8倍。

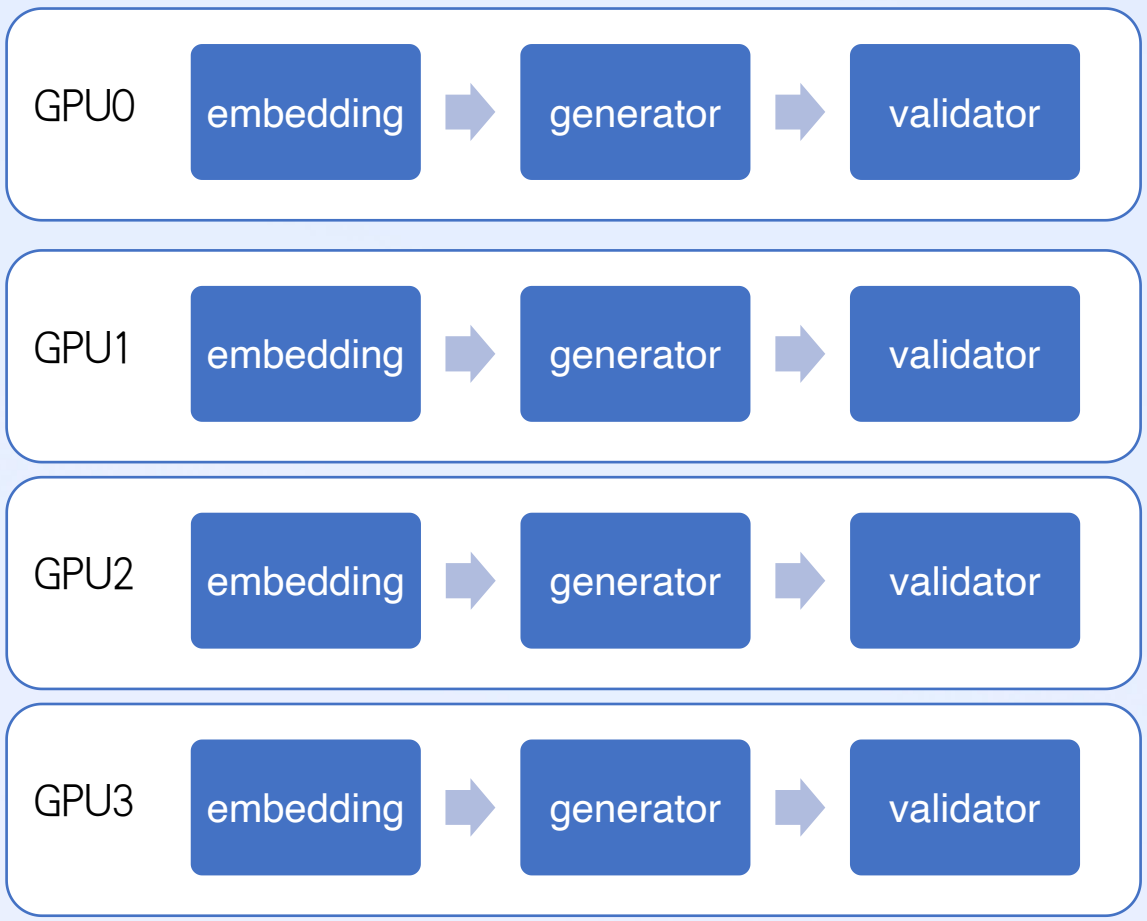
实践案例：第四范式推理加速框架SLX LLM



通常来说，一个完整的大模型商业产品并不仅仅包含一个生成器，而是由一个生成器和若干的小模型组成，以第四范式的模型产品【式说】为例，其中包含了3部分，一个负责前处理的embedding模型，一个生成器generator，一个负责输出的validator。

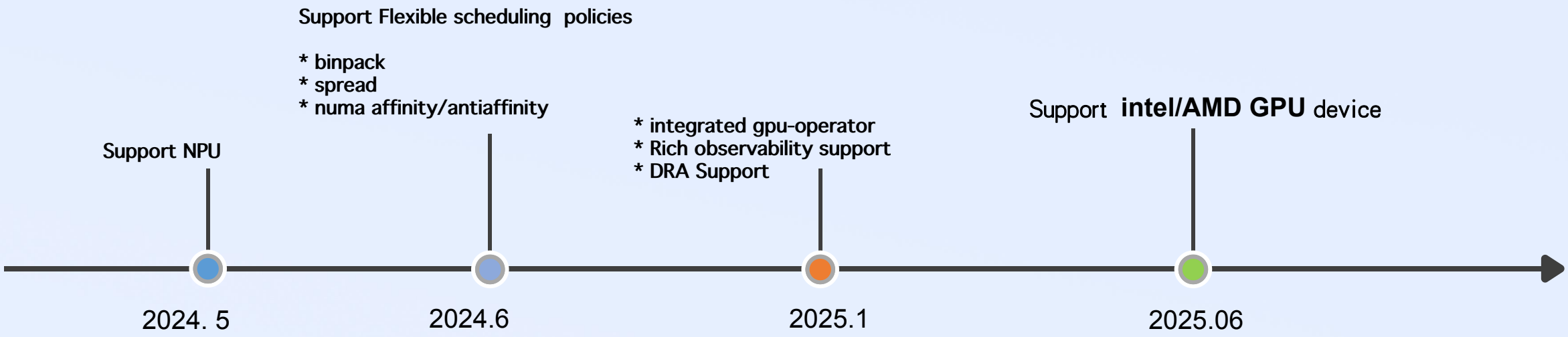
考虑到原生k8s不支持设备复用的问题，最终的部署方案如图所示

实践案例：第四范式推理加速框架SLX LLM

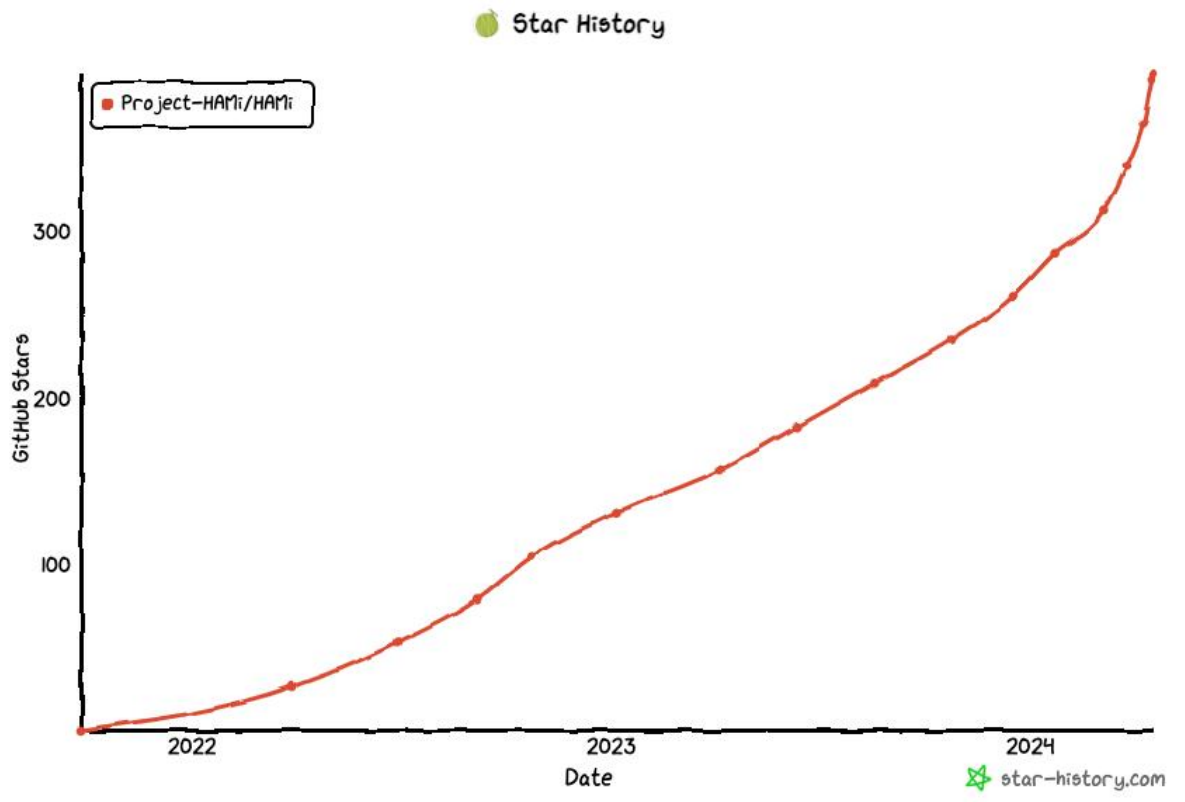


使用哈密瓜可以将这3个组件部署在一张GPU上，因为其中只有一个大模型生成器，embedding和validator均为小模型，以如此部署并不会降低性能，不仅如此，这种部署方式可以在只使用一张GPU的场合部署成功

Roadmap



HAMi Community



- 2022.04 Open Source
- 2024.04 CNCF Landscape project
- Fast growing community
  - 10K+ Downloads
  - 40+ Adopters
  - Aleady Support Nvidia,Cambricon,Hygon,Huawei ASCEND

HAMi Adopters

新网银行

初灵股份 CNCR CHINESE CHURINGA

ICBC 中国工商银行

4Paradigm 第四范式

DaoCloud

中国东信 China-ASEAN Information Harbor

感谢.

AI for everyone.

项目地址：
https://github.com/Project-HAMi/HAMi

商务咨询
business@4paradigm.com

TEL
400-179-1188

北京总部
北京市海淀区清河中街66号
第四范式大厦

上海总部
上海市浦东新区浦东南路1111号
新世纪办公中心15层

深圳总部
深圳市南山区自贸西街151号招商
前海经贸中心一期B座18层1802

新加坡总部
Fourth Paradigm Southeast Asia PTE LTD 1 Fusionopolis Place, #03-20 Galaxis (West Lobby), Singapore, 138522

51CTO WOT

谢 谢 观 看

THANKS

智启新纪
慧创万物
2024