Экспериментальная оценка повышения плотности вычислений на ускорителях Nvidia с применением технологии программно-аппаратного разделения вычислительных сред (Multi-Instance GPU)

А. Малютин, С. Рыкованов, Ю. Шкандыбин

Skoltech

PRECEDENCE RESEARCH

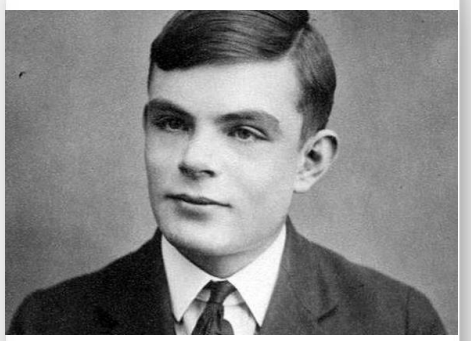
ARTIFICIAL INTELLIGENCE MARKET SIZE, 2021 TO 2030 (USD BILLION)

Во время золотой лихорадки больше всего зарабатывает продавец лопат

Содержание:

1. Что?
2. Как?
3. Зачем?

Истоки

CUDA 8.0	CUDA 9.0	CUDA 10.0	CUDA 11.0
			
2016	2017	2018	2020
PASCAL	VOLTA	TURING	AMPERE
HBM, NVLINK, FP16	HBM, NVLINK, TENSOR CORES, MPS	TENSOR CORES, RT CORES	HBM, NVLINK, TENSOR CORES, PARTITIONING

Сравнение CS, vGPU, MPS, MIG

CUDA 8.0

2016

PASCAL

HBM, NVLINK, FP16

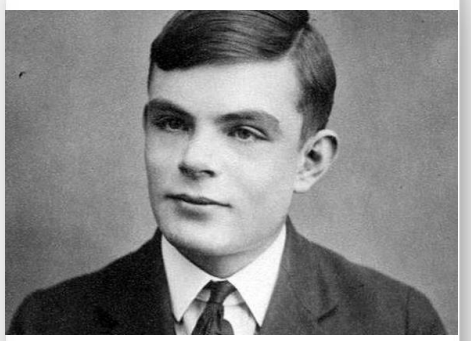
CUDA 9.0

2017

VOLTA

HBM, NVLINK, TENSOR CORES, MPS

CUDA 10.0



2018

TURING

TENSOR CORES, RT CORES

CUDA 11.0

2020

AMPERE

HBM, NVLINK, TENSOR CORES, PARTITIONING

2022

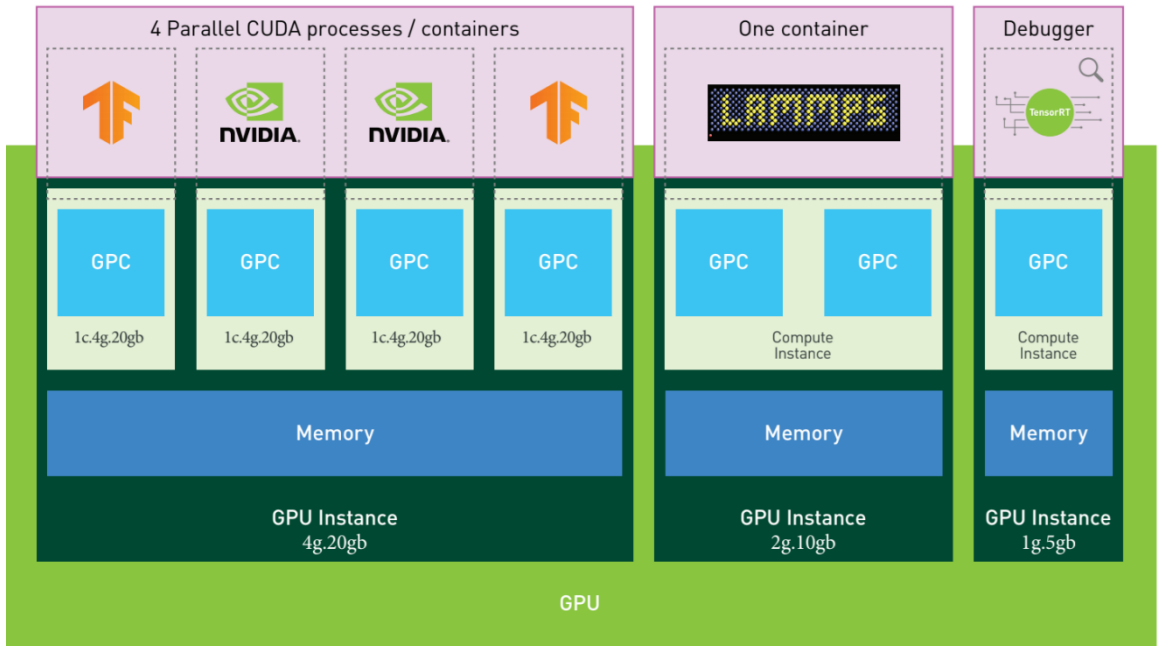
HOPPER

HBM, NVLINK, TENSOR CORES, PARTITIONING

Сравнение CS, vGPU, MPS, MIG

	Parallel work	Address space isolation	SM performance isolation	Memory performance isolation	Error isolation
TRITON (CUDA Streams)	Yes	No	No	No	No
MPS	Yes	Yes	Yes (by percentage, not partitioning)	No	No
vGPU	Yes	Yes (With hypervisor)	Yes (Time-slicing)	Yes	Yes
MIG	Yes	Yes	Yes	Yes	Yes

Особенности MIG на GPU Nvidia A100



До 7 экземпляров GPU в одном A100:
выделенный Steaming Module, память, кэш L2, полоса пропускания для аппаратного QoS

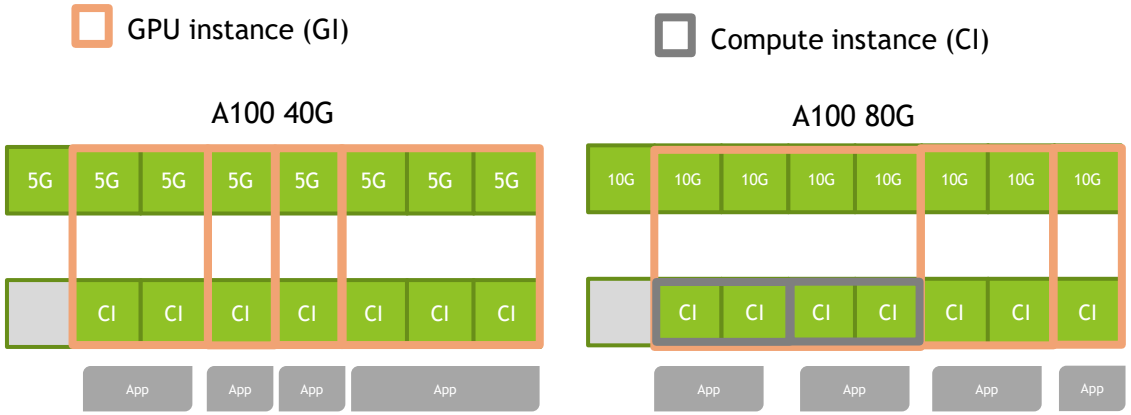
Одновременное выполнение рабочей нагрузки с гарантированным QoS:
все экземпляры MIG работают параллельно с предсказуемой пропускной способностью и задержкой

Динамическое разделение сред:
Возможность изменения конфигурации «на лету»

Гибкость:
Появились GPU instance и Compute instance

Разнообразные среды развертывания:
поддерживается «Bare metal», Docker, Kubernetes, Virtualized Env.

Как делить GPU на MIG устройства



Slice #1	Slice #2	Slice #3	Slice #4	Slice #5	Slice #6	Slice #7
7						
4				2		1
4				1	1	1
2		2		3		
2		1	1	3		
1	1	2		3		
1	1	1	1	3		
3				3		
3				2		1
3				1	1	1
2		2		2		1
2		2		1	1	1
1	1	2		2		1
1	1	2		1	1	1
2		1	1	2		1
2		1	1	1	1	1
1	1	1	1	2		1
1	1	1	1	1	1	1

- 18+ возможных конфигураций
- NVML или NVIDIA-SMI для создания и удаления экземпляра
- Конфигурация может динамически обновляться

НО…
При перезагрузке:
- MIG mode сохраняется
- MIG конфигурация сбрасывается

Ограничения платформы MIG

P2P между графическими процессорами (PCIe или NVLink) не поддерживается.

- Приложения CUDA рассматривают Compute Instance и его родительский GPU Instance как одно устройство CUDA.
- CUDA Inter-Process Communication (IPC) между GPU instances не поддерживается. Поддерживается CUDA IPC для Compute instances.
- Поддерживается отладка CUDA (например, с использованием cuda-gdb) и проверка памяти/race (например, с использованием cuda-memcheck или Compute-sanitizer).
- CUDA MPS поддерживается поверх MIG. Единственное ограничение заключается в том, что максимальное количество клиентов (48) уменьшается пропорционально размеру вычислительного инстанса
- GPUDirect RDMA поддерживается при использовании из GPU instance

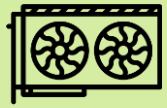
Наш стенд

Виртуальная машина

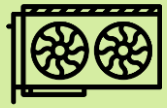
CPU cores: 24
RAM: 192
Disk: SSD

7 x Datasets

GPU0: A100



GPU1: A100



MIG.1

MIG.1

MIG.5

Как работать с MIG

nvidia-smi -i 1 -mig 1

```
NVIDIA-SMI 515.65.01    Driver Version: 515.65.01    CUDA Version: 11.7
+-------------------------------+----------------------+----------------------+
| 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 A100 80G...  Off  | 00000000:0B:00.0 Off |                    0 |
| N/A   31C    P0    41W / 300W |      0MiB / 81920MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  NVIDIA A100 80G...  Off  | 00000000:14:00.0 Off |                   On |
| N/A   40C    P0    75W / 300W |  40294MiB / 81920MiB |     N/A      Default |
|                               |                      |              Enabled |
```

nvidia-smi mig -lgip

```
+--------------------------------------------------------------------------------+
| GPU instance profiles:                                                         |
| GPU   Name             ID    Instances   Memory     P2P    SM    DEC   ENC   |
|                              Free/Total   GiB              CE    JPEG  OFA   |
|================================================================================|
|   0  MIG 1g.5gb        19     0/7         4.75       No     14     0     0   |
|                                                              1     0     0   |
+--------------------------------------------------------------------------------+
|   0  MIG 1g.5gb+me     20     0/1         4.75       No     14     1     0   |
|                                                              1     1     1   |
+--------------------------------------------------------------------------------+
|   0  MIG 2g.10gb       14     0/3         9.75       No     28     1     0   |
|                                                              2     0     0   |
+--------------------------------------------------------------------------------+
|   0  MIG 3g.20gb        9     0/2         19.62      No     42     2     0   |
|                                                              3     0     0   |
+--------------------------------------------------------------------------------+
|   0  MIG 4g.20gb        5     0/1         19.62      No     56     2     0   |
|                                                              4     0     0   |
+--------------------------------------------------------------------------------+
|   0  MIG 7g.40gb        0     0/1         39.50      No     98     5     0   |
|                                                              7     1     1   |
+--------------------------------------------------------------------------------+
```

Создание MIG

nvidia-smi mig -cgi 14,14,14 -C

```
MIG devices:
+------------------+----------------------+-----------+-----------------------+
| GPU  GI  CI  MIG |         Memory-Usage |        Vol|         Shared        |
|      ID  ID  Dev |           BAR1-Usage | SM     Unc| CE  ENC  DEC  OFA  JPG|
|                  |                      |        ECC|                       |
|==================+======================+===========+=======================|
|  1    3   0   0  |   5762MiB / 19968MiB | 28      0 |  2    0    1    0    0 |
|                  |      2MiB / 32767MiB |           |                       |
+------------------+----------------------+-----------+-----------------------+
|  1    4   0   1  |   5914MiB / 19968MiB | 28      0 |  2    0    1    0    0 |
|                  |      2MiB / 32767MiB |           |                       |
+------------------+----------------------+-----------+-----------------------+
|  1    5   0   2  |   5524MiB / 19968MiB | 28      0 |  2    0    1    0    0 |
|                  |      2MiB / 32767MiB |           |                       |
+------------------+----------------------+-----------+-----------------------+
```

nvidia-smi -L

```
GPU 0: NVIDIA A100 80GB PCIe (UUID: GPU-f4a579e9-db9e-3056-2efd-07a2bc89ec0b)
GPU 1: NVIDIA A100 80GB PCIe (UUID: GPU-f8233d9e-8175-3834-2c13-8b8a05ee7356)
  MIG 2g.20gb     Device  0: (UUID: MIG-77820684-17f9-510d-9114-7edc0d9e9664)
  MIG 2g.20gb     Device  1: (UUID: MIG-a682e596-054d-5931-821f-079371c76c5d)
  MIG 2g.20gb     Device  2: (UUID: MIG-b1f2db14-fd90-5096-aea4-2937f1d83626)
```

Удаление MIG

nvidia-smi mig -dci -ci 0,1,2 -gi 1

Именования в MIG

NVIDIA_VISIBLE_DEVICES supports the following formats to specify MIG devices:

- **MIG-<GPU-UUID>/<GPU instance ID>/<compute instance ID>** when using R450 and R460 drivers or MIG-<UUID> starting with R470 drivers.
- **<GPUDeviceIndex>:<MIGDeviceIndex>**

```
NVIDIA_VISIBLE_DEVICES =MIG-GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0/1/0
NVIDIA_VISIBLE_DEVICES="0:0“
```

CUDA_VISIBLE_DEVICES supports the following formats to specify MIG devices:

- **MIG-<UUID>** with drivers >= R470
- **MIG-<GPU-UUID>/<GPU instance ID>/<compute instance ID>** with drivers R450 and R460

```
CUDA_VISIBLE_DEVICES=MIG-77820684-17f9-510d-9114-7edc0d9e9664
CUDA _VISIBLE_DEVICES =MIG-GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0/1/0
```

В MIG нет привычных устройств:

```
/dev
├── nvidiactl
├── nvidia-modeset
├── nvidia-uvm
├── nvidia-uvm-tools
├── nvidia-nvswitchctl
├── nvidia0
└── nvidia1
```

Как работать с MIG в Docker

 # Для Docker versions < 19.03

```
$ sudo docker run --runtime=nvidia \
    -e NVIDIA_VISIBLE_DEVICES="0:0" \
    nvidia/cuda nvidia-smi -L
GPU 0: A100-SXM4-40GB (UUID: GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0)
  MIG 3g.20gb Device 0: (UUID: MIG-GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0/1/0)
```

 # Для Docker versions >= 19.03

```
$ sudo docker run --gpus '"device=0:0"' \
    nvidia/cuda nvidia-smi -L
GPU 0: A100-SXM4-40GB (UUID: GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0)
  MIG 3g.20gb Device 0: (UUID: MIG-GPU-e86cb44c-6756-fd30-cd4a-1e6da3caf9b0/1/0)
```

 # Как запусить задачу

```
$ sudo docker run --gpus '"device=0:1"' \
    nvcr.io/nvidia/pytorch:20.11-py3 \
    /bin/bash -c 'cd /opt/pytorch/examples/upstream/mnist && python main.py'
```

Особенности работы с MIG

```
$ docker run \
 --gpus ‘”device=0:0,0:1”’ \
 nvidia/cuda:11.0-base nvidia-smi -L
GPU 0: A100-SXM4-40GB (UUID: GPU-2ceff3df-31b3-caf2-eace-a494b4b7926b)
  MIG 3g.20gb Device 0: (UUID: MIG-GPU-2ceff3df-31b3-caf2-eacea494b4b7926b/1/0)
  MIG 3g.20gb Device 1: (UUID: MIG-GPU-2ceff3df-31b3-caf2-eacea494b4b7926b/2/0)
```

```
$ docker run \
--gpus ‘”device=MIG-GPU-2ceff3df-31b3-caf2-eace-a494b4b7926b/1/0” ‘ \
nvidia/cuda:11.0-base nvidia-smi -L
GPU 0: A100-SXM4-40GB (UUID: GPU-2ceff3df-31b3-caf2-eace-a494b4b7926b)
  MIG 3g.20gb Device 0: (UUID: MIG-GPU-2ceff3df-31b3-caf2-eacea494b4b7926b/1/0)
```

UUID после перезагрузки изменяется

Как мониторить MIG

NVIDIA **Data Center GPU Manager**

Накопительный мониторинг

```
dcgmi group --create GPU1
dcgmi group -g 2 -a 0
dcgmi stats -g 2 –enable
dcgmi stats -g 2 -s RUN
dcgmi stats -g 2 -j RUN
```

создаем группу
добавляем карту 0
включаем подсчет статистики
запускаем сбор
смотрим данные

```
GROUPS
3 groups found.
+-----------------+----------------------------------------------------------+
| Groups          |                                                          |
| -> 0            |                                                          |
|    -> Group ID  | 0                                                        |
|    -> Group Name| DCGM_ALL_SUPPORTED_GPUS                                  |
|    -> Entities  | GPU 0, GPU 1                                             |
| -> 1            |                                                          |
|    -> Group ID  | 1                                                        |
|    -> Group Name| DCGM_ALL_SUPPORTED_NVSWITCHES                            |
|    -> Entities  | None                                                     |
| -> 2            |                                                          |
|    -> Group ID  | 2                                                        |
|    -> Group Name| GPU1                                                     |
|    -> Entities  | GPU 1                                                    |
+-----------------+----------------------------------------------------------+

Summary
+------------------------------------+-----------------------------------------+
|-----  Execution Stats  ------------+-----------------------------------------|
| Start Time                         | Sun Sep 11 11:53:40 2022                |
| End Time                           | Sun Sep 11 11:55:35 2022                |
| Total Execution Time (sec)         | 115.1                                   |
| No. of Processes                   | 0                                       |
|-----  Performance Stats  ----------+-----------------------------------------|
| Energy Consumed (Joules)           | 2698                                    |
| Power Usage (Watts)                | Avg: 45.0107, Max: N/A, Min: N/A        |
| Max GPU Memory Used (bytes)        | 0                                       |
| Clocks and PCIe Performance        | Available per GPU in verbose mode       |
|-----  Event Stats  -----------------+-----------------------------------------|
```

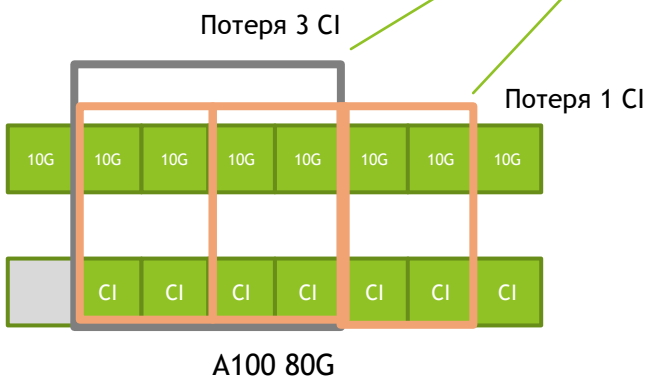
Потоковый монторинг

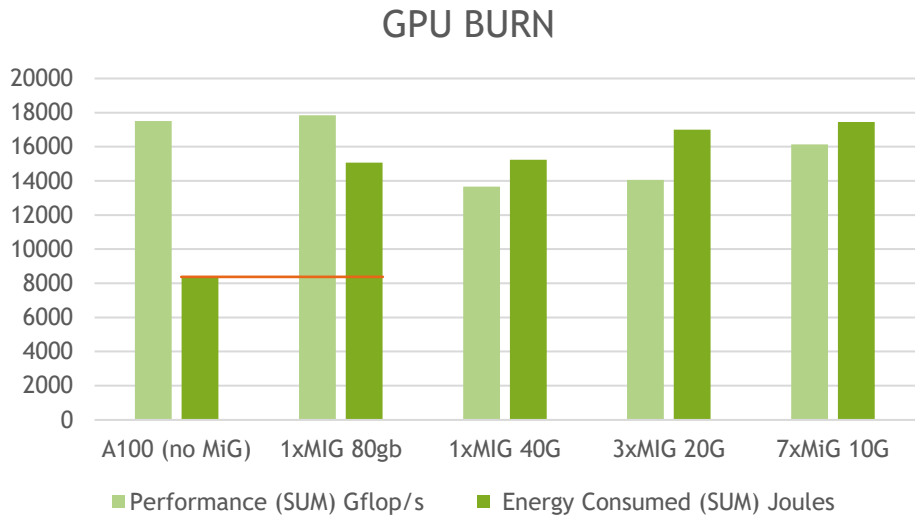
```
dcgmi dmon -e 1001,1002,1004,1005,1009,1010,1011,1012,150,155,110,111
```

```
# Entity  GRACT  SMACT  TENSO  DRAMA          PCITX          PCIRX          NVLTX          NVLRX  TMPTR     POWER  SACLK  MACLK
    Id                                                                                          C         W
  GPU 0  0.931  0.777  0.175  0.496      175899291     1532954951     1547634958     1553333956     52   323.689   1410   1215
  GPU 1  0.948  0.780  0.173  0.496      172945598     1507859117     1522127704     1522126460     50   213.963   1410   1215
  GPU 2  0.952  0.778  0.175  0.493      178507418     1557783818     1572668487     1572504828     48   359.610   1410   1215
  GPU 3  0.962  0.793  0.178  0.503      164054321     1428701446     1327745638     1327396166     52   226.107   1410   1215
  GPU 4  0.960  0.786  0.179  0.499      163908021     1430858946     1288201051     1287639531     64   392.270   1410   1215
  GPU 5  0.952  0.797  0.182  0.506      182644334     1599554874     1235853101     1233988554     62   341.524   1410   1215
  GPU 6  0.966  0.817  0.200  0.508      132741767     1148660264     1129637355     1127111684     64   258.063   1410   1215
  GPU 7  0.999  0.867  0.325  0.451        8908656       34245363              0              0     67   380.955   1410   1215
  GPU 0  0.950  0.793  0.179  0.505      162992146     1418772939     1429455794     1435422194     54   304.839   1410   1215
  GPU 1  0.954  0.793  0.179  0.505      162944796     1418947251     1430185344     1430185344     52   201.105   1410   1215
  GPU 2  0.959  0.795  0.179  0.505      162966713     1419469072     1430752928     1430665363     53   372.072   1410   1215
  GPU 3  0.962  0.796  0.179  0.505      162992814     1418956709     1430003326     1429872315     56   195.564   1410   1215
  GPU 4  0.960  0.792  0.179  0.505      162681409     1418483393     1431427800     1430779751     66   400.533   1410   1215
  GPU 5  0.948  0.789  0.179  0.506      162794813     1419172557     1435095820     1431911846     65   355.586   1410   1215
  GPU 6  0.957  0.794  0.179  0.505      162844371     1418843705     1439494242     1434260115     67   292.070   1410   1215
  GPU 7  0.958  0.797  0.180  0.506      163341225     1422783028     1440327549     1443650284     70   384.132   1410   1215
  GPU 0  0.949  0.793  0.179  0.505      163030005     1419242144     1431636763     1440810765     55   237.659   1410   1215
  GPU 1  0.954  0.793  0.179  0.505      162773503     1418681427     1431965210     1431965210     53   184.241   1410   1215
  GPU 2  0.957  0.795  0.179  0.506      162881890     1419208015     1432242919     1432242919     53   366.301   1410   1215
  GPU 3  0.959  0.797  0.179  0.506      163018599     1419626682     1432350659     1432350659     56   225.281   1410   1215
  GPU 4  0.957  0.792  0.180  0.506      162612068     1418432187     1433020167     1432423976     66   396.763   1410   1215
  GPU 5  0.949  0.790  0.179  0.506      162527136     1417457784     1436326347     1431757595     65   278.087   1410   1215
  GPU 6  0.956  0.794  0.179  0.505      162874341     1419286866     1446253398     1437670082     68   330.311   1410   1215
  GPU 7  0.957  0.797  0.181  0.506      162804092     1419116905     1441752283     1446325537     71   402.675   1410   1215
  GPU 0  0.951  0.796  0.180  0.507      163049364     1419855949     1431014322     1437445669     55   178.524   1410   1215
  GPU 1  0.951  0.795  0.180  0.507      162991159     1420276463     1431002457     1431002457     52   237.659   1410   1215
  GPU 2  0.958  0.797  0.180  0.507      162890127     1419226321     1430767359     1430767359     53   366.837   1410   1215
  GPU 3  0.960  0.797  0.179  0.506      162951049     1419130718     1430934017     1430371626     56   320.427   1410   1215
  GPU 4  0.957  0.794  0.181  0.506      162632607     1418105919     1431840408     1430961031     67   387.907   1410   1215
  GPU 5  0.948  0.791  0.180  0.507      162692025     1418504737     1435486856     1431624416     66   202.787   1410   1215
  GPU 6  0.958  0.795  0.180  0.507      162784709     1418531037     1441021899     1435473584     68   383.275   1410   1215
  GPU 7  0.955  0.798  0.182  0.507      162806309     1418604281     1436151038     1440574380     71   408.494   1410   1215
  GPU 0  0.954  0.795  0.180  0.506      162990387     1418928837     1429538441     1434605453     55   231.030   1410   1215
  GPU 1  0.953  0.794  0.179  0.506      162726406     1418134181     1429561807     1429561807     53   327.149   1410   1215
  GPU 2  0.957  0.795  0.179  0.506      162749917     1418119150     1429417688     1429417688     54   318.781   1410   1215
  GPU 3  0.960  0.798  0.179  0.506      162980413     1418831034     1429747257     1429747257     56   366.668   1410   1215
  GPU 4  0.958  0.794  0.181  0.507      162550382     1417626219     1430946788     1429692646     67   355.460   1410   1215
  GPU 5  0.948  0.790  0.179  0.506      162619862     1417919747     1434448412     1431062748     65   225.026   1410   1215
  GPU 6  0.958  0.796  0.180  0.506      162713169     1418220725     1438159351     1434345897     68   394.858   1410   1215
  GPU 7  0.956  0.797  0.181  0.506      162676691     1417962711     1434673417     1438058602     72   403.272   1410   1215
```

Результаты графики GPUBURN

	Performance (SUM) Gflop/s	Energy Consumed (SUM) Joules
A100 (no MiG)	17502	8414
1xMIG 80gb	17836	15070
1xMIG 40G	13656	15232
3xMIG 20G	14053	17002
7xMiG 10G	16130	17455

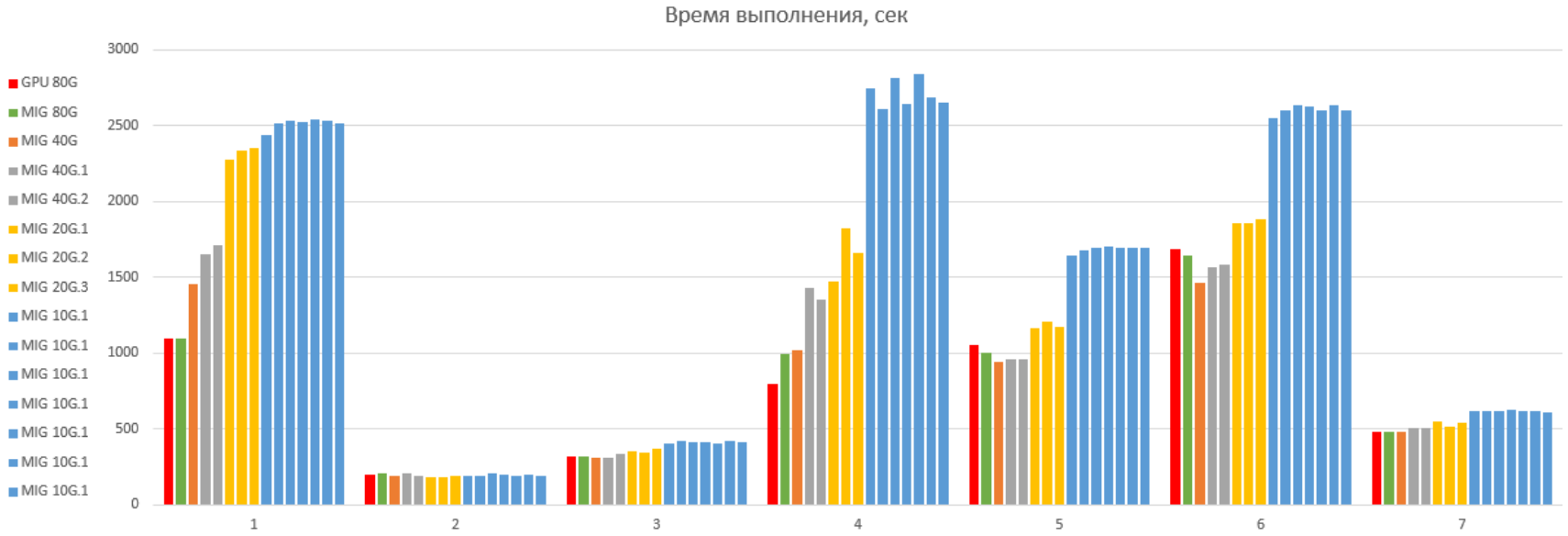


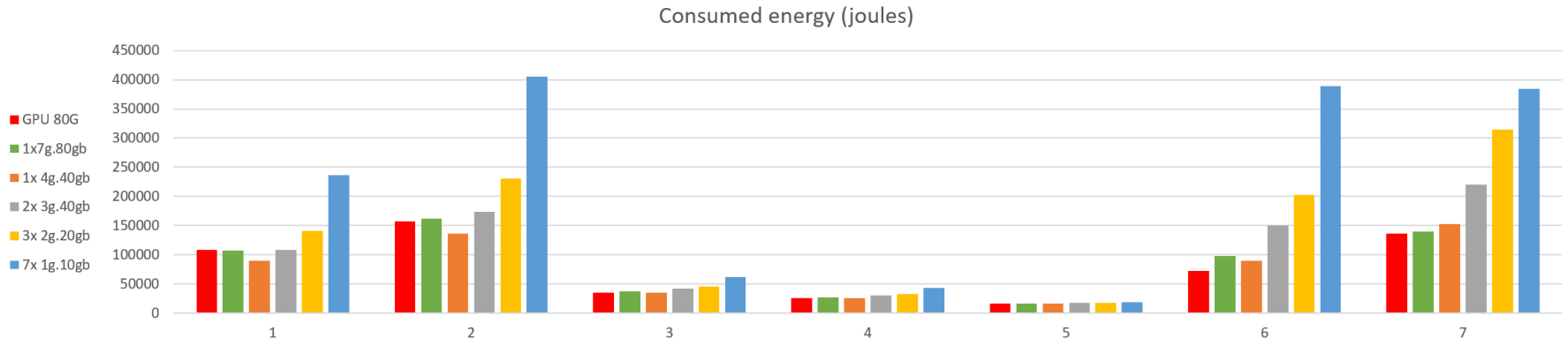


- Падение производительности не превышает 22%
- Существенно увеличивается потребление электропитания

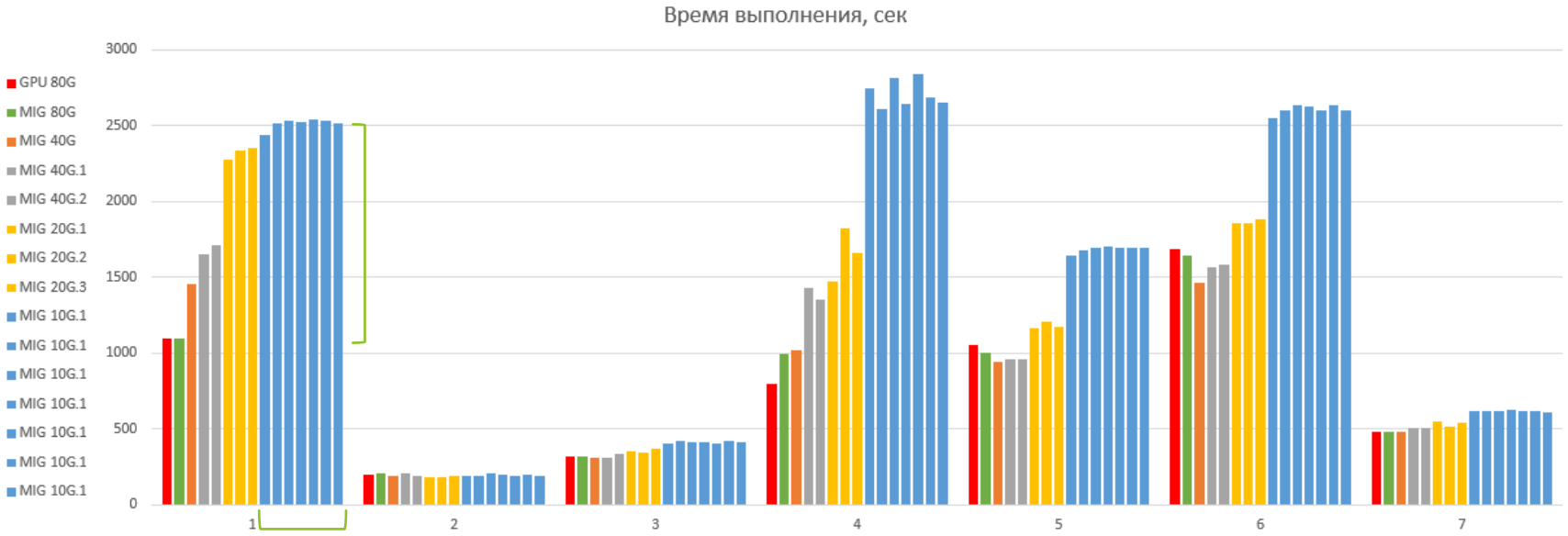
https://github.com/wilicc/gpu-burn

Результаты расчетов для 7 дата сетов



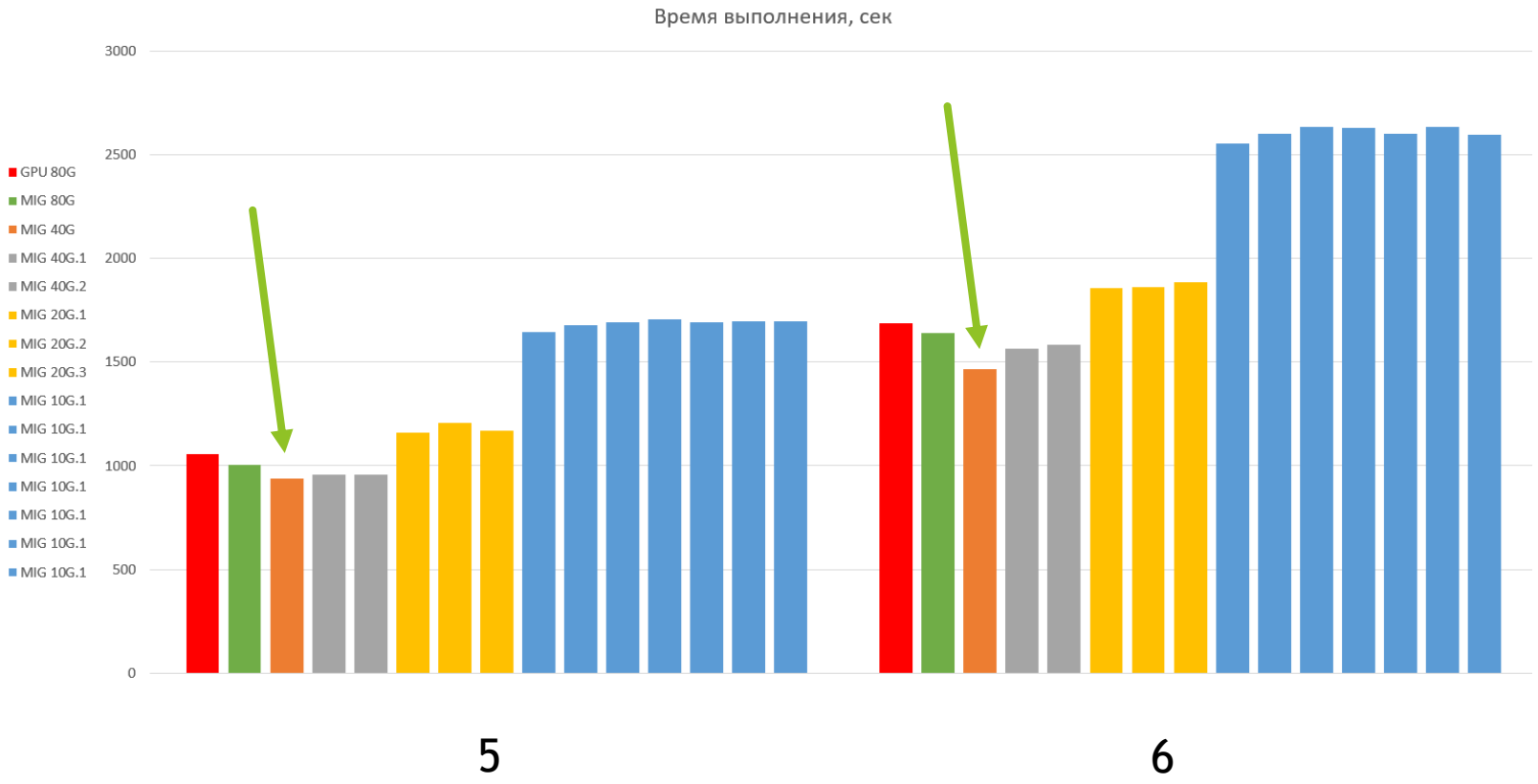


Результаты графики 7 data sets



Тип	Среднее уменьшение производительности на один экземпляр MIG	Средняя суммарная производительность на группу экземпляров
MIG 2x40G	**80%**	**159%**
MIG 3x20G	**72%**	**215%**
MIG 7x10G	**58%**	**408%**

Результаты и комментарии



Дата сеты лучше помещаются в меньший объем памяти

Выводы

Используя технологию MIG можно кратно увеличить производительность вычислений

до 7x

Асинхронная загрузка данных в память экземпляров MIG выравнивает загрузку CPU и снижает нагрузку на подсистему ввода-вывода

Using Multi-Instance GPU for Efficient Operation of Multi-Tenant GPU Clusters

K8s, Slurm и взгляд в будущее

Поддержка MIG Kubernetes
https://docs.nvidia.com/datacenter/cloud-native/kubernetes/mig-k8s.html

Getting Started

Install Kubernetes

As a first step, ensure that you have a Kubernetes deployment set up with a control plane and nodes joined to the cluster. Follow the Install Kubernetes guide for getting started with setting up a Kubernetes cluster.

Configuration Strategy

TBD.

MIG доступен в новой версии Slurm
https://slurm.schedmd.com/gres.html#MIG_Management

Beginning in version 21.08, Slurm now supports NVIDIA Multi-Instance GPU (MIG) devices. This feature allows some newer NVIDIA GPUs (like the A100) to split up a GPU into up to seven separate, isolated GPU instances. Slurm can treat these MIG instances as individual GPUs, complete with cgroup isolation and task binding.

DCGM-Exporter
https://docs.nvidia.com/datacenter/cloud-native/gpu-telemetry/dcgm-exporter.html
https://grafana.com/grafana/dashboards/16640-nvidia-mig-dcgm-exporter-dashboard/

Позволяет пользователям собирать показатели графического процессора и понимать поведение рабочей нагрузки или отслеживать графические процессоры в кластерах.

Спасибо за внимание

Email: an.maliutin@skoltech.ru

Telegram: @malutinanton

Skoltech

Skolkovo Institute of Science and Technology