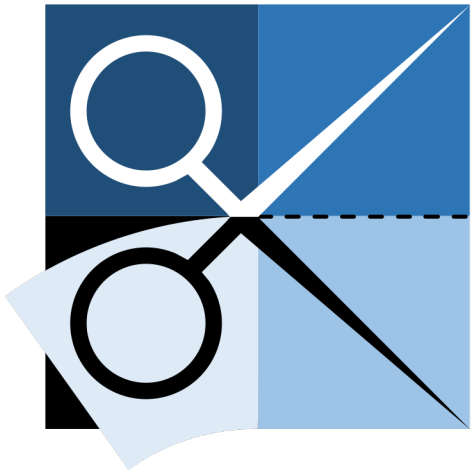
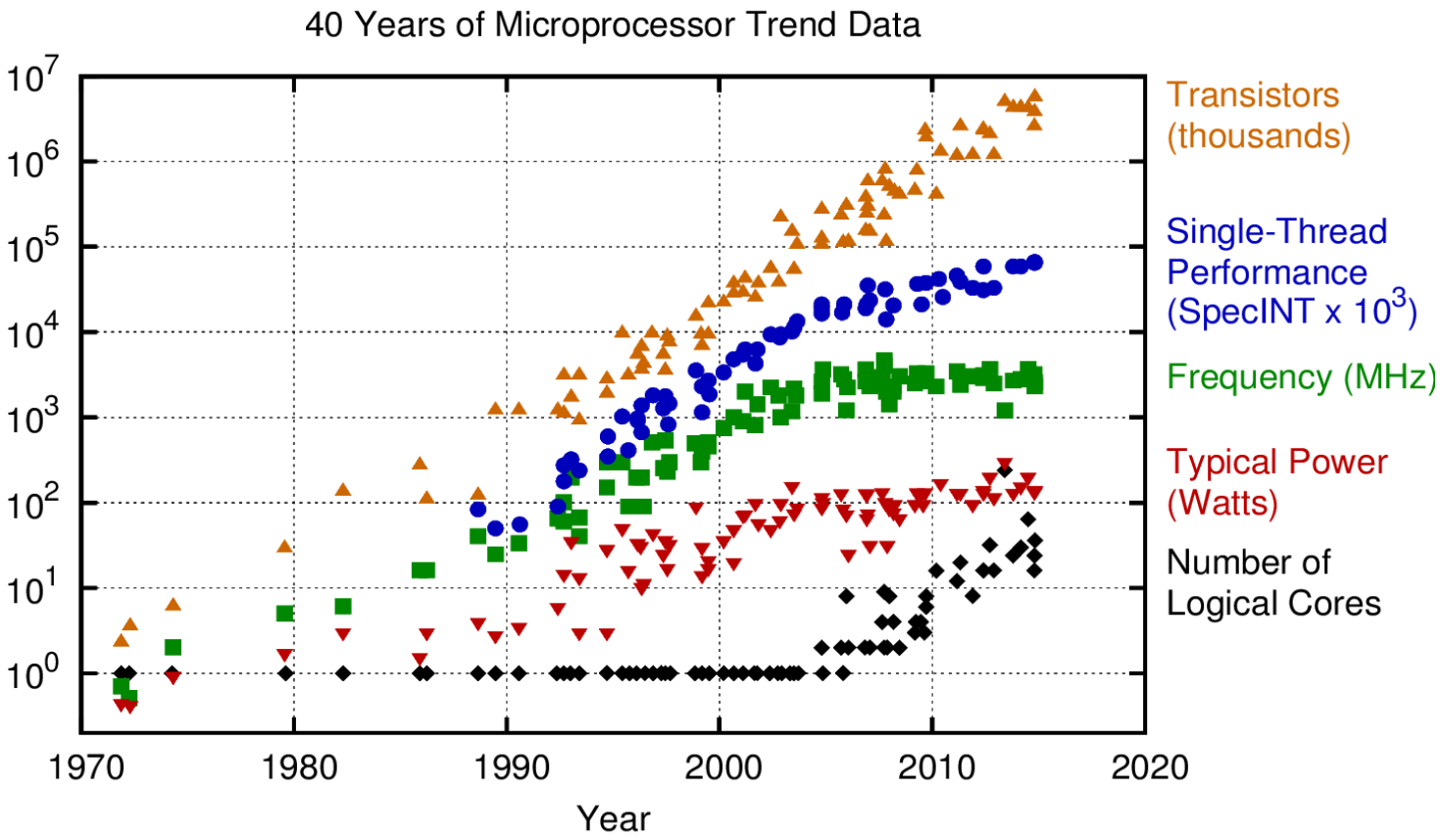


OpenForBC

the definitive partitionable GPU interface

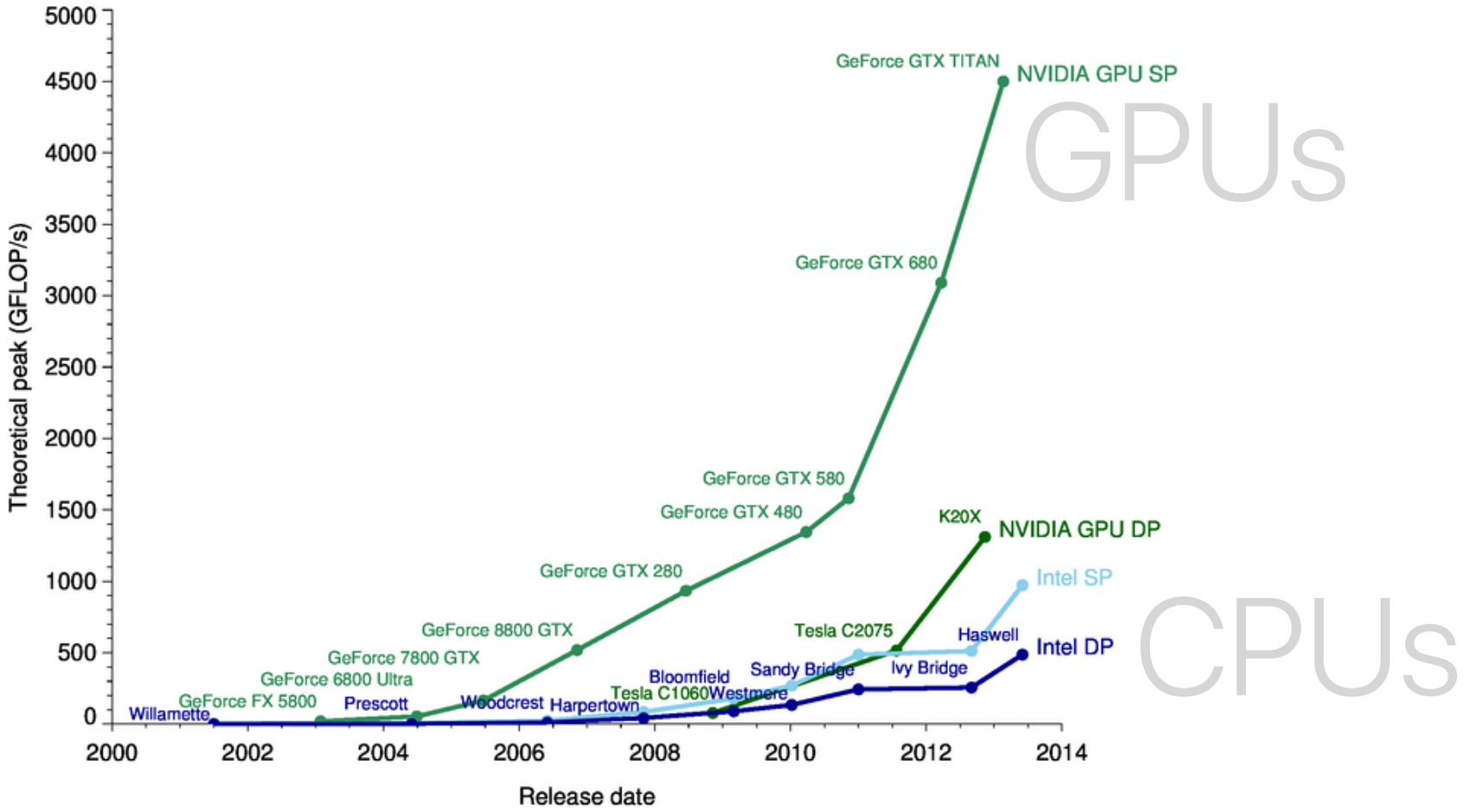
Federica Legger e Gabriele Gaetano Fronzé - INFN Torino

RATIONALE

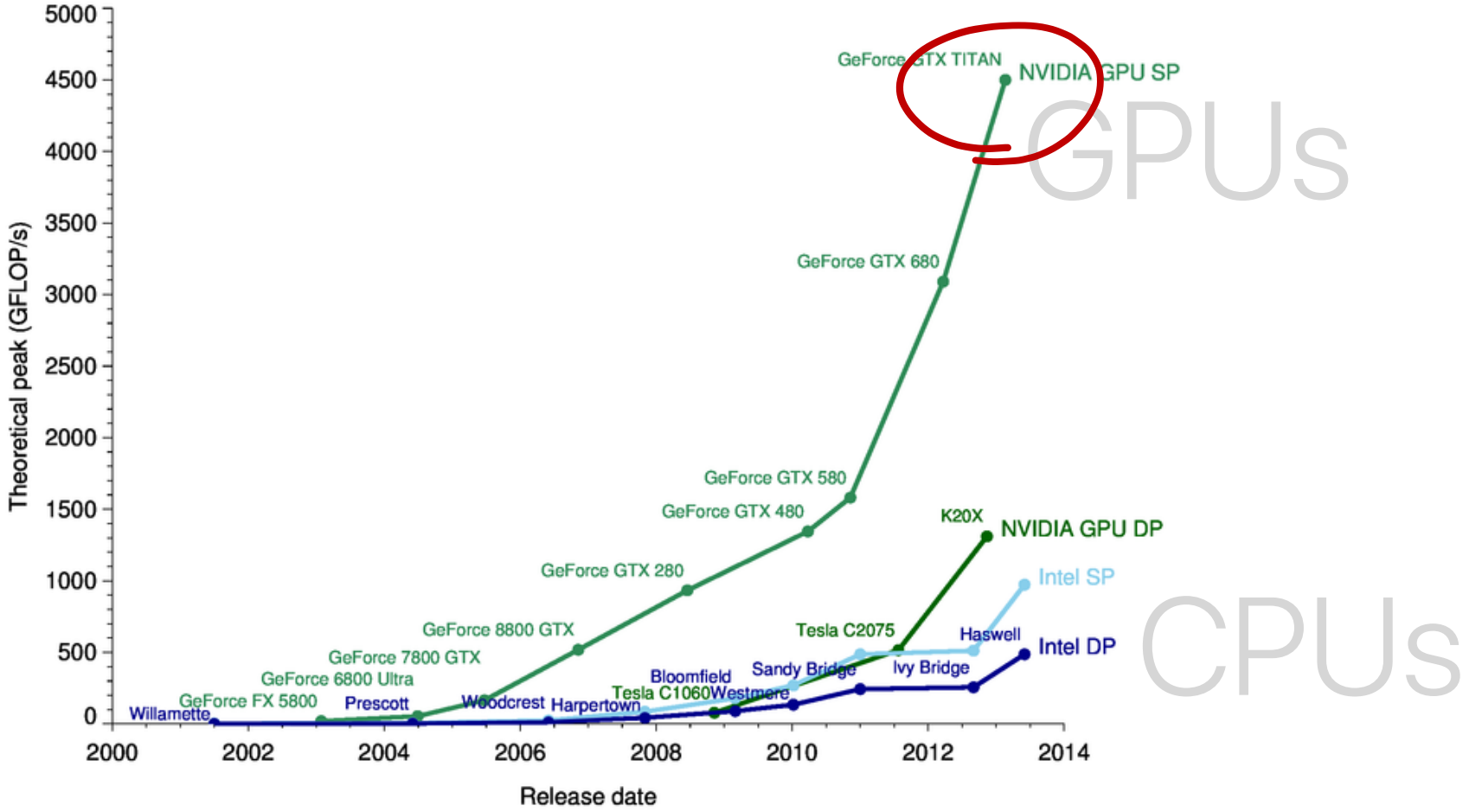


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

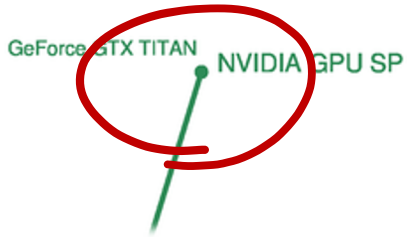
RATIONALE


Theoretical peak (GFLOP/s)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Release date
GeForce GTX TITAN
NVIDIA GPU SP
GeForce GTX 680
GeForce GTX 580
GeForce GTX 480
GeForce GTX 280
GeForce 8800 GTX
GeForce 7800 GTX
GeForce 6800 Ultra
GeForce FX 5800
Willamette
Prescott
Woodcrest
Harpertown
Bloomfield
Tesla C1060
Westmere
Sandy Bridge
Ivy Bridge
Haswell
Tesla C2075
K20X
NVIDIA GPU DP
Intel SP
Intel DP
GPUs
CPUs

RATIONALE


Theoretical peak (GFLOP/s)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
2000
2002
2004
2006
2008
2010
2012
2014
Release date
GeForce GTX TITAN
NVIDIA GPU SP
GPUs
GeForce GTX 680
GeForce GTX 580
GeForce GTX 480
GeForce GTX 280
GeForce 8800 GTX
GeForce 7800 GTX
GeForce 6800 Ultra
GeForce FX 5800
Willamette
Prescott
Woodcrest
Harpertown
Bloomfield
Tesla C1060
Westmere
Sandy Bridge
Ivy Bridge
Haswell
Tesla C2075
K20X
NVIDIA GPU DP
Intel SP
Intel DP
CPUs

RATIONALE



Is that **too much**?

Can we **exploit** that power?

Is that **efficient**?

RATIONALE

Is that **too much**?

Can we **exploit** that power?

Is that **efficient**?

STATE OF ART

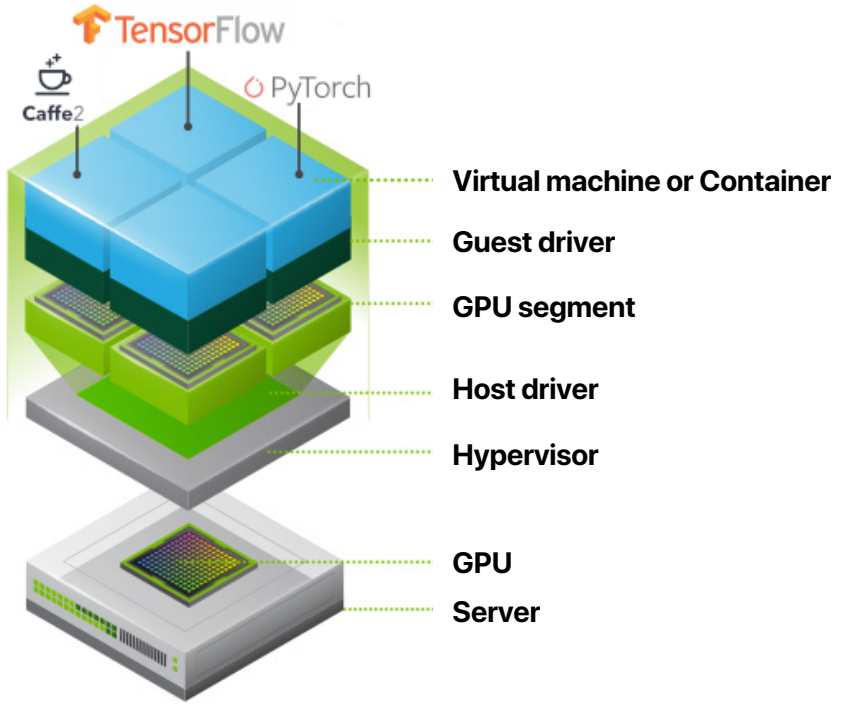
USING VIRTUAL MACHINES
DIVIDE ET IMPERA!

- Smaller GPUs are easier to use **efficiently**
- Developing on a small GPU is **accessible**
- Some tasks are by design **inefficient on huge GPUs**
- Large GPUs can be seen as a **set of smaller ones**

STATE OF ART

USING VIRTUAL MACHINES

DIVIDE ET IMPERA!

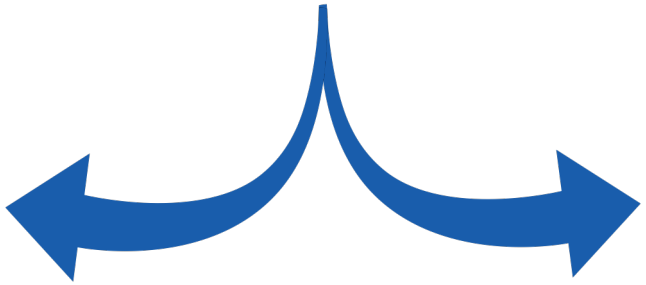


AUTEM...

STATE OF ART

PCI EXPRESS®

Single Root I/O Virtualization



NVIDIA®

GRID, vGPU, MIG

closed source

AMD

MxGPU

open source

Everyone is adopting its own "**standard**" derived from the only real one

STATE OF ART

Single Root I/O Virtualization

 Same underlying **standard**

 Same technical **operations** and **procedures**

 **Different** top level **semantics**

STATE OF ART

PCI EXPRESS®

Single Root I/O Virtualization

Proprietary “standards” only partially implemented in Linux

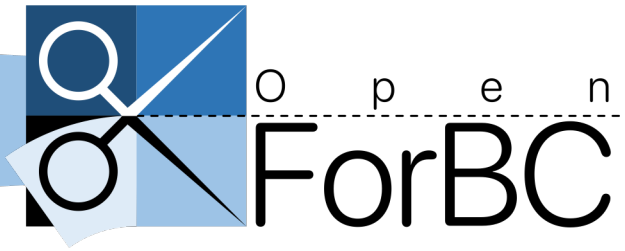
OPEN ForBC

Single Root I/O Virtualization

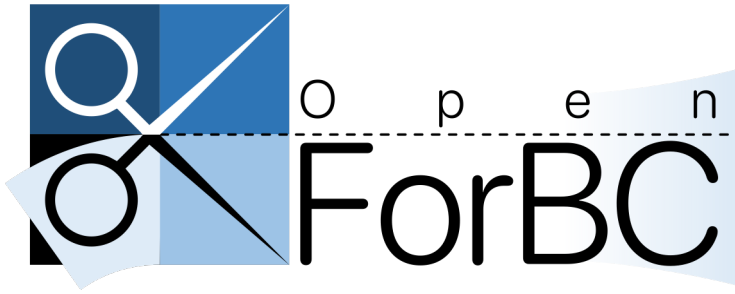
GRID, vGPU, MIG

MxGPU

Open For Better Computing



- **Uniform interface** for GPU partitioning
- Common toolset for **future new technologies**
- **No vendor specificity left**
- Improved **Linux Compatibility**

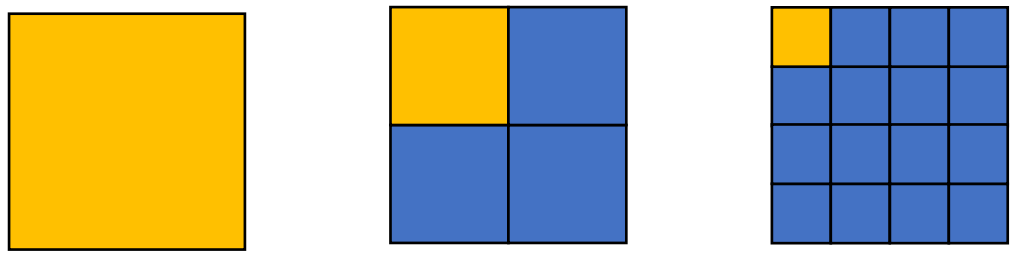



```
get_available_modes

set_mode

get_current_mode

get_instance
```

Open ForBC

get_available_modes

Lists the available virtual GPU profiles, with plenty of information about peak performances and memory size.

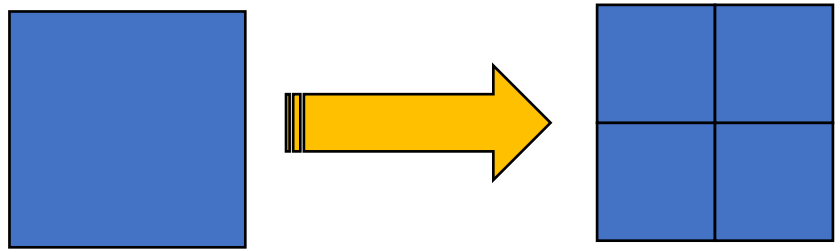


1 vGPU | 4 vGPUs | 16 vGPUs

Open ForBC

set_mode

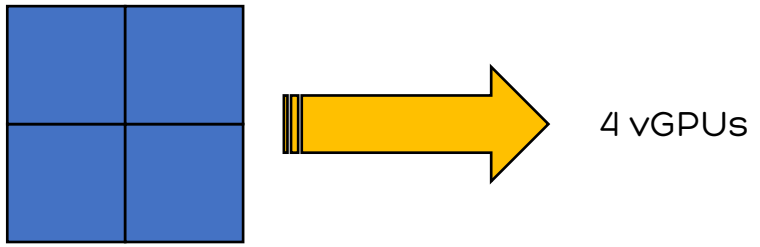
Applyes one of the available profiles performing the required procedures and ensuring not to disrupt existing instances



Open ForBC

get_current_mode

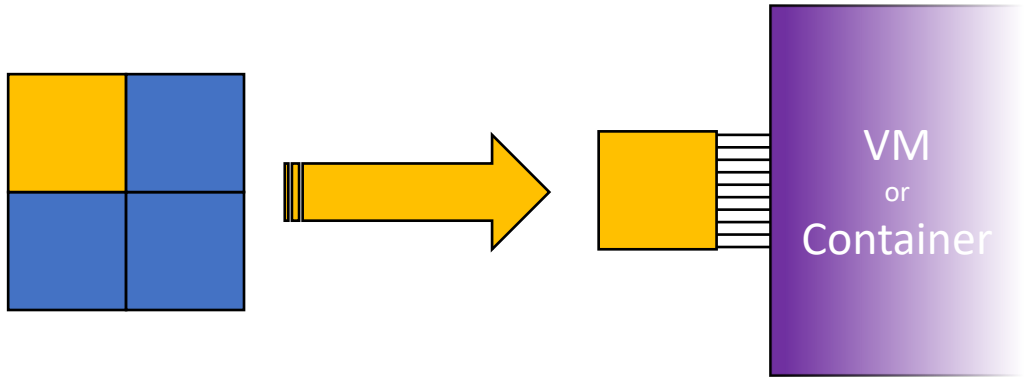
Returns the mode currently set up on the GPU.



Open ForBC

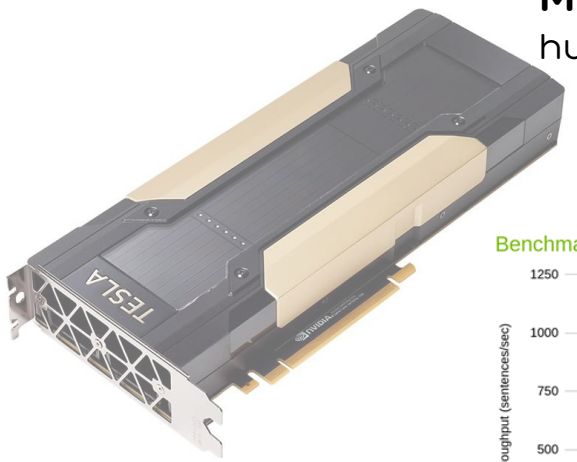
get_instance

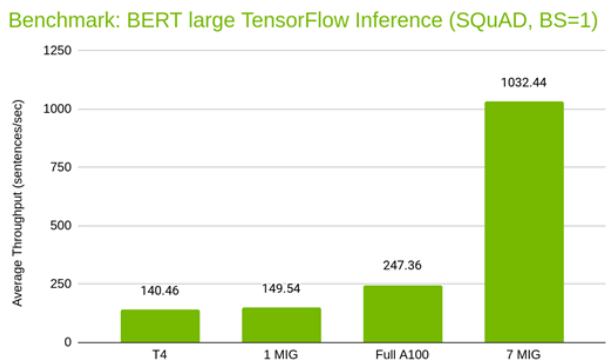
Retrieves the information needed to instanciate a new VM or container attached to the vGPU instance.

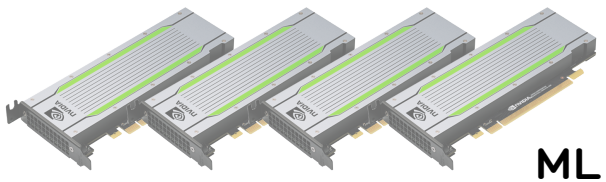


Hardware choice:
Nvidia V100 32GB GPU

ML training:
huge GPU with lots of memory



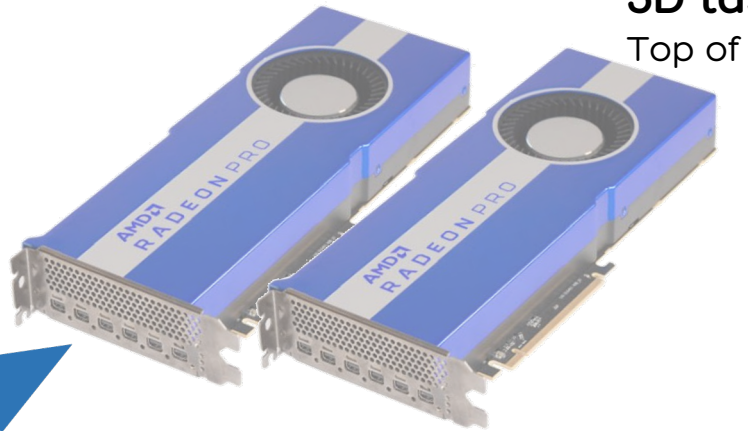




ML inference:
smaller GPUs for higher throughput

OPEN ForBC

Hardware choice:
AMD Radeon PRO V340



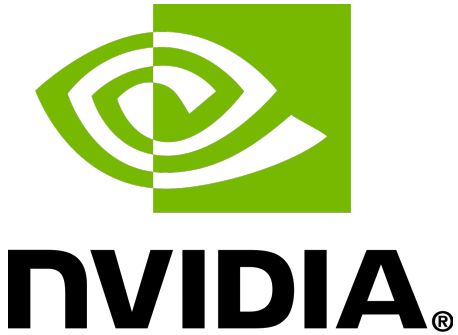
3D tasks:
Top of the line vGPU instances

Up to 128 users per server
Down to 1/6 power consuption

2D tasks:
Up to 32 vGPUs (32 users!)

OPEN ForBC

Partner tecnici e *sample provider*:



Federica Legger: 0.2 FTE

Gabriele Gaetano Fronzé: 0.2 FTE

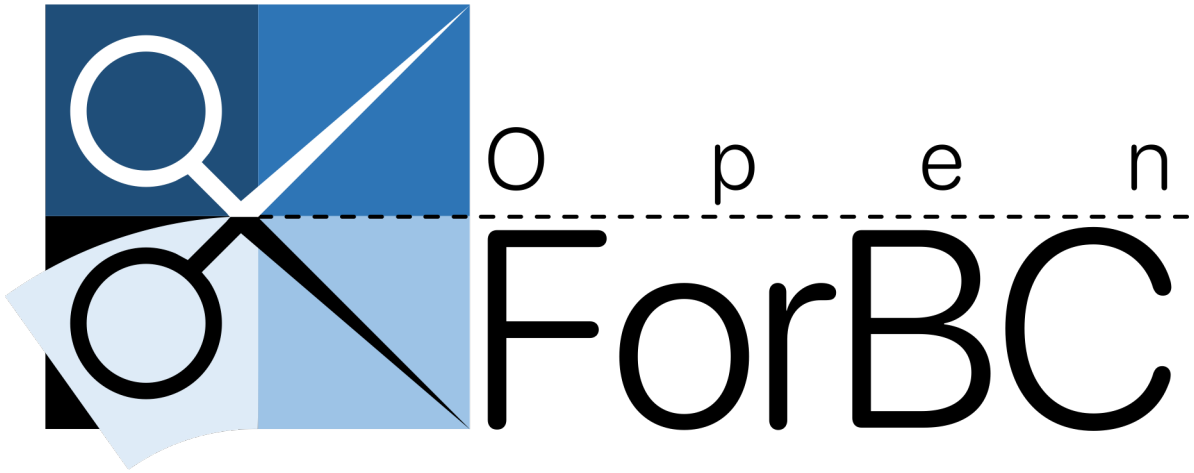
AdR 1 anno: 1.0 FTE

OPEN ForBC

Open ForBC introduces a smarter way to use partitionable GPUs.

Optimizing your hardware deployment for your tasks is few clicks away.

Never depend on manifacturer choices: with Open ForBC learn one and use forever.



Thank you for your attention!

Federica Legger e Gabriele Gaetano Fronzé - INFN Torino