GPU Acceleration in HPC

Derek Bouius, Product Manager – AI Computing
Radeon Technology Group, AMD
FLEXIBLE GPU SERVER INFRASTRUCTURE

Multiple GPUs per Server (e.g. Radeon Instinct™ MI25)

Configurable software environment to target many different workloads

Dual CPU, lots of Memory, high speed Storage and Interconnect
MANY SOFTWARE DEPLOYMENT OPTIONS

Bare Metal

Virtualization

Containers

- App1
- App2
- App3

- Bins/Libs
- OS

- Server
- GPU-0
- GPU-1
- ...
- GPU-n

- Guest OS
- Hypervisor
- Host OS

- Server
- GPU-0
- GPU-1
- ...
- GPU-n

- Container
- App1
- App2
- App3

- Bins/Libs
- OS

- Server
- GPU-0
- GPU-1
- ...
- GPU-n
CONTAINER AND CLUSTER MANAGEMENT

kubernetes

https://github.com/RadeonOpenCompute/k8s-device-plugin
FLEXIBLE CONTAINERS ENABLED BY HCC2

https://github.com/ROCm-Developer-Tools/hcc2

hcc2: Heterogeneous Compiler Collection (Version 2)
Experimental PROTOTYPE that is intended to support multiple programming models including

- OpenMP 4.5+
- OpenCL
- HIP
- Cuda

- Supports offloading to multiple GPU acceleration targets (multi-target)
- Supports different host platforms such as AMD64, PPC64LE, and AARCH64
Thank You

derek.bouius@amd.com