GPU Acceleration in HPC

Derek Bouius, Product Manager –AI Computing

Radeon Technology Group, AMD

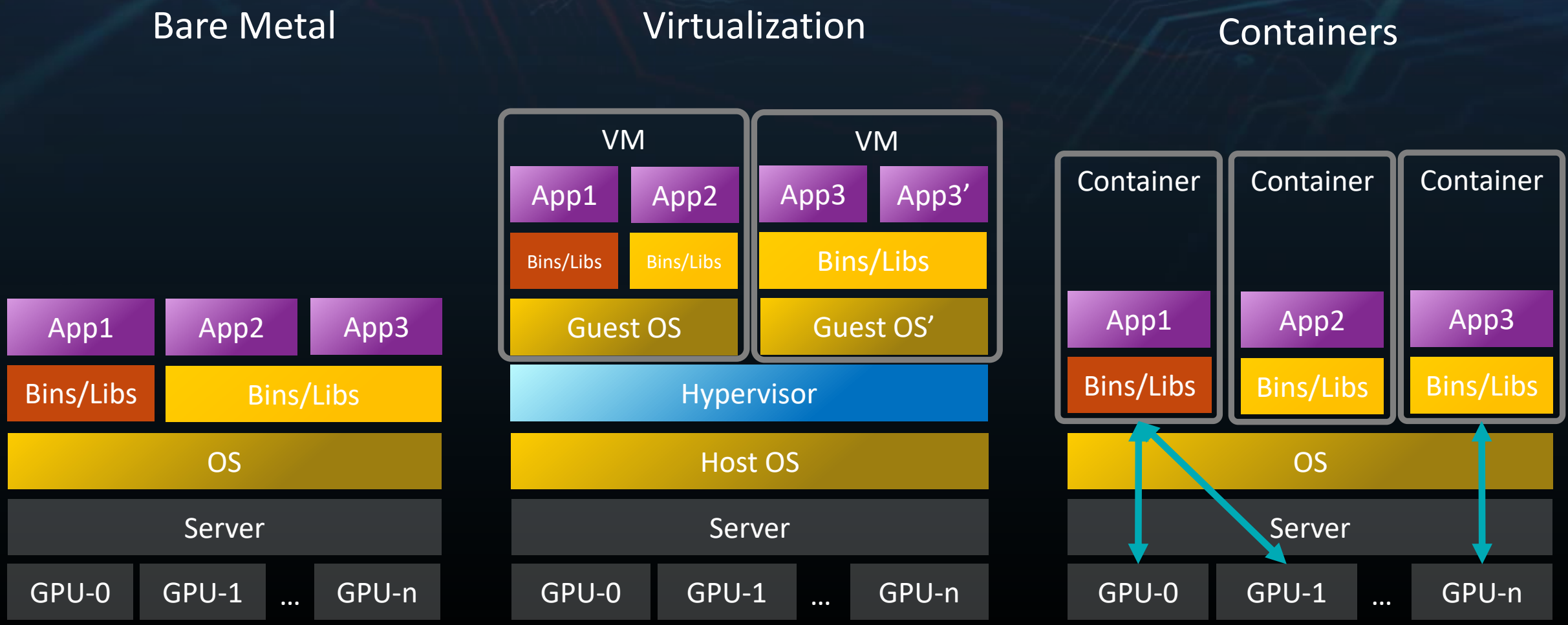
FLEXIBLE GPU SERVER INFRASTRUCTURE

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

Dual CPU, lots of Memory, high speed Storage and Interconnect

Configurable software environment to target many different workloads

MANY SOFTWARE DEPLOYMENT OPTIONS



CONTAINER AND CLUSTER MANAGEMENT

kubernetes

docker

Concepts

- Overview
- Compute, Storage, and Networking Extensions
- Kubernetes Architecture
- Extending Kubernetes
  - Extending your Kubernetes Cluster
  - Extending the Kubernetes API
  - Compute, Storage, and Networking Extensions
    - Network Plugins
    - Device Plugins
    - Service Catalog
- Containers
- Workloads
- Configuration
- Services, Load Balancing, and Networking
- Storage
- Policies

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