TFDx - NetApp

Andy Banta

Enhanced Data Center Workloads
Not Your father's Oldsmobile

First-generation DCs
Resource-constrained. As long as something was limiting, things don't get done.
Late first-generation DCs - standalone hosts with apps
2nd-generation DCs - multiple workloads on hosts,
2nd-generation DCs with virtualisation

Compute, memory, throughput, and capacity
The DC is full of boxes
Remember "Monster VMs", some exceptions with speciality PCIe add-in cards (but that ties the VM to a host / hosts)

*Persistent Memory
In-memory DBs (been around since before persistent memory)
In-memory DBs with resiliency (SAP HANA, REDIS, MemSQL)

128 - 512GB on one stick of memory, slot form factors. You can get really large density and split it between persistent and dynamic

Now you can slice up this persistent memory and share with VMs, and you provision them out as persistent datastore

*Microvisors
There's no such thing as a full stack developer

Serverless & SaaS
Containers
VMs
Hypervisors / Microvisor
Bare Metal

Why are we nesting all of these?

Cloud, on-premises, and hybrid - depends on ownership, location, and mindset

Why Microvisors?

Hypervisor	Microvisor	Containers
Secure Isolation between workloads & tenants	Secure isolation between workloads & tenants	Poor isolation between workloads / tenants on same kernel
Excellent hardware isolation & control	Just enough hardware isolation & control	Minimal hardware isolation & control
"Slow" boot up time	Fast enough boot up time	"Fast" startup time
upports many "heavy" OSes	Supports one lightweight OS	No OS, sharing kernel

Microvisors are free - like a puppy is free

- Still a VM with its own OS
- has minimal modern hardware virtualisation, i.e. no floppy, just NVMe
- No BIOS and much simpler code
- only supports modern Linux

*Specialised cores, on demand
Graphics applications for GPUs
- oil and gas
- CAD / CAM (eg F1 cars)
- medical industry (eg MRIs)
- media and entertainment
- gaming industry

Graphic virtualisation for VDI
- 8K, 4K, HD, etc all remotely displayed

Inference / AI
CUDA cores - graphics
Tensor core - AI

AI / ML / Deep learning

How artificial is AI?
Perception and categorisation of the world around us
Contextualisation of relationships between things
Prediction for cause and effect
Planning and decision making based on external and internal factors

The reality?
Better than humans at some perception and categorisation. But they suck at the rest of it.

NVIDIA Inference Server

Who uses this stuff, anyway?
Oil & Gas
Facial recognition on drivers licences - identify identity theft, other fraud
Google Earth (formerly Keyhole Systems)
Railroads

How it's used with VMware?
Direct - 1 user, 1 GPU
Shared Direct - many users, 1 GPU
Shared - many users, many GPUs

*HPC
HPC using NVMeoF

Workloads
- banking - low-latency transactions
- fluid dynamics - lots of data being processed quickly in a parallel stream
- medical and nuclear research

"NVMe is grafting flesh back on to the skeleton of fibre channel" - FC is dead.
It's out there for high margin vendors
The more you can run on commodity hardware, the better off you are.

NVMe over fabric as the next InfiniBand

The future?
vSCSI -> vNVMe
iSCSI, FC, NFS -> NVMeoF

We've gone from everyone sitting in their own rows in the DC, to workloads and VMs taking advantage of all types of technologies

Memory-enhanced FlexPod
NVIDIA M10 and T4 GPU nodes
VMware integration with NVMeoF